UNIVERSITY OF CALIFORNIA,
IRVINE

Essays on Missing Data Models, BLP Contraction Mappings, and MCMC Estimation

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Economics

by

Phillip Li

Dissertation Committee:
Professor David Brownstone, Chair
Professor Ivan Jeliazkov
Professor Dale Poirier

2012

DEDICATION

To my parents, Huan and Susan Li,
for their unconditional support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am grateful to Professors David Brownstone (committee chair), Ivan Jeliazkov, and Dale Poirier for their guidance throughout my graduate studies. In particular, I would like to thank Professor Brownstone for supporting me financially through multiple research assistantships and for always giving me useful advice. His generosity and positive influence have been invaluable. I also benefited tremendously from Ivan Jeliazkov. He is approachable and incredibly encouraging. I am thankful to him for guiding me through my first two publications. Finally, I thank Professor Poirier for recruiting me to UC Irvine, entrusting me to TA his graduate course for the last four years, and being instrumental in the establishment of the joint Statistics degree.

I would also like to acknowledge the love and support from Erica Tom. She is my best friend and my foundation. Without her, I would not have succeeded.

Funding provided by the School of Social Sciences, Department of Economics, UC Irvine, and UC MultiCampus Research Program in Sustainable Transport: Technology, Mobility, Infrastructure are acknowledged. I would also like to thank all my classmates for their support and helpful discussions.

The copyright for Chapter 1 belongs to Elsevier. It is published as “Estimation of sample selection models with two selection mechanisms” in *Computational Statistics and Data Analysis*, Vol 55, pp 1099-1108 (2011). The copyright for Chapter 2 (with Arshad Rahman) belongs to Emerald Publishing Limited. It is published as a chapter entitled “Bayesian analysis of multivariate sample selection models using Gaussian copulas” in *Advances in Econometrics: Missing-data Methods* in 2011.

CURRICULUM VITAE

Phillip Li

EDUCATION

Doctor of Philosophy in Economics **2012**
University of California, Irvine *Irvine, California*

Master of Science in Statistics **2011**
University of California, Irvine *Irvine, California*

Master of Arts in Economics **2009**
University of California, Irvine *Irvine, California*

Bachelor of Arts in Economics **2006**
University of California, Berkeley *Berkeley, California*

PUBLICATIONS

"Estimation of sample selection models with two selection mechanisms" **2011**
Computational Statistics and Data Analysis

"Bayesian analysis of multivariate sample selection models using Gaussian copulas" **2011**
Advances in Econometrics: Missing-data Methods

PROFESSIONAL EXPERIENCE

Graduate Research Assistant **2008—2009, 2010**
University of California, Irvine *Irvine, California*

Teaching Assistant **2007—2012**
University of California, Irvine *Irvine, California*

SELECTED HONORS AND AWARDS

David Brownstone Award for Best Econometrics Paper **2011**
University of California, Irvine

Outstanding Graduate Student Award **2010**
University of California, Irvine

Outstanding Economics Teaching Assistant Award **2008, 2010**
University of California, Irvine

School of Social Science Merit Fellowship **2007**
University of California, Irvine

Graduated with Honors and Distinction in Economics **2007**
University of California, Berkeley

ABSTRACT OF THE DISSERTATION

Essays on Missing Data Models, BLP Contraction Mappings, and MCMC Estimation

By

Phillip Li

Doctor of Philosophy in Economics

University of California, Irvine, 2012

Professor David Brownstone, Chair

My dissertation is composed of four chapters that focus on missing data models, BLP contraction mappings, and Markov chain Monte Carlo estimation.

The first chapter focuses on estimating sample selection models with two incidentally truncated outcomes and two corresponding selection mechanisms. The method of estimation is an extension of the Markov chain Monte Carlo (MCMC) sampling algorithm from Chib (2007) and Chib et al. (2009). Contrary to conventional data augmentation strategies for dealing with missing data, the proposed algorithm augments the posterior with only a small subset of the total missing data caused by sample selection. This results in improved convergence of the MCMC chain and decreased storage costs, while maintaining tractability in the sampling densities. The methods are applied to estimate the effects of residential density on vehicle miles traveled and vehicle holdings in California. The empirical results suggest that residential density has a small economic impact on vehicle usage and holdings. In addition, the results show that changes to vehicle holdings from increased residential density are more sensitive for less fuel-efficient vehicles than for fuel-efficient vehicles on average.

The second chapter considers the estimation of a multivariate sample selection model

with p pairs of selection and outcome variables. A unique feature of this model is that the variables can be discrete or continuous with any parametric distribution, allowing a large class of multivariate models to be accommodated. For example, the model may involve any combination of variables that are continuous, binary, ordered, or censored. Although the joint distribution can be difficult to specify, a multivariate Gaussian copula function is used to link the marginal distributions together and handle the multivariate dependence. The proposed estimation approach relies on the MCMC-based techniques from Lee (2010) and Pitt et al. (2006) and adapts the methods from the preceding authors to a missing data setting. An important aspect of the estimation algorithm, in the same spirit as the algorithm from the first chapter, is that it does not require simulation of the missing outcomes. This has been shown to improve the mixing of the Markov chain. The methods are applied to both simulated and real data.

The third paper analyzes a discrete choice model where the observed outcome is not the exact alternative chosen by a decision maker but rather the broad group of alternatives in which the chosen alternative belongs to. This model is designed for situations where the choice behavior at a lower level is of interest but only higher level data are available (e.g. analyzing households' choices for vehicles at the make-model-trim level but only choice data at the make-model level are observed). I show that the parameters in the proposed model are locally identified, but for certain configurations of the data, they are weakly identified. Methods to incorporate additional information into the problem are discussed, and both maximum likelihood and Bayesian estimation methods are explored.

The last chapter proposes improvements to the contraction mappings used in the context of multinomial logit models. The contraction mapping algorithm proposed in Berry et al. (1995) is slow to converge and is a major burden to implement in applied

work. While it is relatively quick to converge for a single run of the algorithm, it is computationally expensive when repeated evaluations are needed, particularly when the algorithm is embedded into maximum likelihood, generalized method of moments, or Bayesian Markov chain Monte Carlo estimation routines. To alleviate this problem, I explore four simple modifications of the contraction mapping to improve its rate of convergence. Importantly, the modifications can be incorporated into existing code with minimal effort. In a simulation study, I demonstrate that the new algorithms require significantly fewer iterations to converge to the unique vector of fixed points than the original specification. The best algorithm results in an 80-fold improvement.

Chapter 1

Estimation of sample selection models with two selection mechanisms

1.1 Introduction

The seminal sample selection model of Heckman (1976, 1979b) has generated a vast amount of theoretical and empirical research across a variety of disciplines. Sample selection, also referred to as incidental truncation, occurs when a dependent variable of interest is non-randomly missing for a subset of the sample as a result of a separate selection variable. A well-known application involves market wages (the outcome of interest) and labor force participation (the selection variable), in which wages are missing for individuals who are not participating in the labor force. Consequently, the remaining observations available to the researcher are non-random and do not represent the population of interest. As a result, estimation based only on this selected sample may lead to specification errors. This problem is ubiquitous in economics and disciplines that use observational data, therefore estimation techniques that address this issue are of substantial interest.

The conventional sample selection model with a single selection mechanism and its variants have been extensively estimated. Common classical estimation methods are developed and discussed in Amemiya (1984), Gronau (1973), Heckman (1976, 1979b), and Wooldridge (1998, 2002), while semiparametric estimation and a variety of extensions are discussed in Heckman (1990), Manski (1989), and Newey et al. (1990). Extensions in the direction of multiple selection mechanisms are discussed in Shonkwiler and Yen (1999), Yen (2005), and Poirier (1980), where the two former articles discuss a model similar to that presented here, and the latter discusses observability of a single binary outcome as a result of two binary selection variables. The preceding procedures generally involve two classes of estimators: 1) two-step estimators that are consistent, asymptotically normal, but inefficient, and 2) maximum likelihood estimators that depend on evaluations of integrals. Puhani (2000) studies the practical performance of both classes of estimators using a Monte Carlo framework and criticizes their small sample properties. Alternatively, Bayesian estimation results in finite sample inference and avoids direct evaluations of integrals. Recent developments with one selection mechanism include Chib et al. (2009), Greenberg (2007), and van Hasselt (2009); extensions such as semiparametric estimation, endogeneity, and multiple outcome types are also discussed.

The model being analyzed contains a correlated system of equations with two continuous dependent variables of interest, each with an incidental truncation problem, and two corresponding selection variables. A major difference between this model and much previous work is that there are two incidentally truncated outcomes being considered instead of one, resulting in four possible combinations of missing data for any observational unit and a complex pattern of missing data across the entire sample. The main contribution of this article is in the extension of the Markov chain Monte Carlo (MCMC) algorithm with "minimal data augmentation" to accommodate the nature of missing data for the model being analyzed. The minimal data augmentation technique first appeared in Chib (2007) to estimate the Bayesian version of the Roy model and was later extended in Chib et al. (2009) to estimate a semiparametric model with endogeneity and sample selection. This paper proposes an algorithm that only involves a minimal subset of the total missing data in the sampling scheme, resulting in improved convergence of the Markov chain and decreased storage costs, while maintaining tractability of the sampling densities without the complete data. The sampling densities are easy to draw from and result in samples that are close to iid for many parameters. A simulation study is included to study the performance of the algorithm.

The methods are applied to study the effects of residential density on vehicle miles traveled and vehicle holdings in California. A careful analysis is needed since data for vehicle miles traveled is only observable for households that own vehicles. The resulting estimation results will supplement the current literature and be informative for policy decisions.

1.2 Sample Selection Model

The model is given by

$$\begin{aligned}
y_{i,1} &= x'_{i,1}\beta_1 + \epsilon_{i,1}, && (1.1)\\
y_{i,2} &= x'_{i,2}\beta_2 + \epsilon_{i,2}, && (1.2)\\
y^*_{i,3} &= x'_{i,3}\beta_3 + \epsilon_{i,3}, && (1.3)\\
y^*_{i,4} &= x'_{i,4}\beta_4 + \epsilon_{i,4}, && (1.4)\\
y_{i,j} &= t_j \quad \text{if} \quad \alpha_{t_j-1,j} < y^*_{i,j} \leq \alpha_{t_j,j}, && (1.5)\\
\delta_{t_j,j} &= ln\left\{\frac{\alpha_{t_j,j} - \alpha_{t_j-1,j}}{1-\alpha_{t_j,j}}\right\}, && (1.6)
\end{aligned}$$

for observational units $i = 1, \ldots, N$, equations $j = 3, 4$, ordered categories $t_j = 1, \ldots, T_j$, ordered cutpoints $\alpha_{0,j} = -\infty < \alpha_{1,j} = 0 < \alpha_{2,j} = 1 < \alpha_{3,j} < \ldots < \alpha_{T_j-1,j} < \alpha_{T_j,j} = +\infty$, and transformed cutpoints $\delta_{t_j,j}$ for $t_j = 3, \ldots, T_j - 1$. The cutpoint restrictions are discussed in Section 1.3.3. The continuous dependent variables of interest are $y_{i,1}$ and $y_{i,2}$. Due to sample selection, their observability depends on the values of two ordered selection variables, $y_{i,3}$ and $y_{i,4}$ from (1.5), respectively. Following Albert and Chib (1993), the ordered variables are modeled in a threshold-crossing framework with the latent variables $y^*_{i,3}$ and $y^*_{i,4}$ according to equations (1.3) through (1.5). In addition, a re-parameterization of the ordered cutpoints according to equation (1.6) is performed to remove the ordering constraints (Chen and Dey, 2000). The row vector $x'_{i,j}$ and conformable column vector β_j are the exogenous covariates and corresponding regression coefficients, respectively. The vector of error terms $(\epsilon_{i,1}, \epsilon_{i,2}, \epsilon_{i,3}, \epsilon_{i,4})'$ is distributed independent multivariate normal, $\mathcal{N}(0, \Omega)$, where Ω is an unrestricted covariance matrix. This normality assumption for the

Variables	N_1	N_2	N_3	N_4
$y_{i,1}$	✓	○	✓	○
$y_{i,2}$	✓	✓	○	○
$y_{i,3}$	✓	✓	✓	✓
$y_{i,4}$	✓	✓	✓	✓

Table 1.1: Variable observability. The symbols ○ and ✓ denote whether the variable is missing or observed in the sample partition, respectively.

error terms results in ordered probit models for equations (1.3) through (1.5).

A key feature of the model is the inclusion of two incidentally truncated outcomes, which results in four cases of observability. For any observational unit i, only one of the following vectors is observed

$$(y_{i,1}, y_{i,2}, y_{i,3}, y_{i,4})', \quad (y_{i,2}, y_{i,3}, y_{i,4})', \quad (y_{i,1}, y_{i,3}, y_{i,4})', \quad (y_{i,3}, y_{i,4})', \tag{1.7}$$

where $y_{i,1}$ and $y_{i,2}$ are missing if and only if $y_{i,3}$ and $y_{i,4}$ are in known, application-specific categories γ and λ, respectively. In the context of the empirical application, the mileage driven with trucks and cars ($y_{i,1}$ and $y_{i,2}$) are missing when the number of trucks and cars owned by the household ($y_{i,3}$ and $y_{i,4}$) equal zero, expressed as $y_{i,3} = \gamma = 0$ and $y_{i,4} = \lambda = 0$. The rules involving $y_{i,3}$ and $y_{i,4}$ that affect the observability are known as the selection mechanisms. To be specific about where incidental truncation occurs, let N_r ($r = 1, \ldots, 4$) denote partitions of the sample set that correspond to the four aforementioned cases of observability in (1.7). In addition, let n_r denote their sizes such that $\sum_{r=1}^{4} n_r = N$. Using this notation, the variable $y_{i,1}$ is only observed for units in $N_1 \cup N_3$, and $y_{i,2}$ is only observed for units in $N_1 \cup N_2$, as illustrated in Table 1.1. Other quantities such as the ordered variables and explanatory variables are always observed.

1.3 Estimation

The proposed estimation algorithm uses MCMC methods with minimal data augmentation (MDA) based on Chib (2007) and Chib et al. (2009). The idea, motivation, and implementation of MDA are described in Section 1.3.1. Section 1.3.2 provides the data-augmented likelihood, priors, and data-augmented posterior. Section 1.3.3 presents the sampling algorithm in detail.

1.3.1 Minimal Data Augmentation (MDA)

The aim of MDA is to augment the posterior with the least amount of missing outcomes possible while keeping the densities of interest tractable for sampling. By introducing all the latent and missing data along the lines of Tanner and Wong (1987), many complex econometric models can be estimated as linear regression models with Gibbs or Metropolis-Hastings sampling (see Chapter 14 of Koop et al. (2007) for many examples). This approach is often desirable since given the "complete" data, the full conditional densities for $\tilde{\beta}$, Ω, and other quantities are in standard forms (Chib and Greenberg, 1995). However, as noted in Chib et al. (2009), such a "naive" approach would degrade the mixing of the Markov chains and increase computation time. This problem is especially intensified when the quantity of missing outcomes due to the selection mechanism is large or when the model contains a sizable number of unknown parameters. Even if these impediments are disregarded, sample selection makes simulating the missing outcomes difficult as influential covariates may also be missing as a result of sample selection. For these reasons, it is generally desirable to minimize the amount of missing outcomes involved in the algorithm.

The proposed algorithm only augments the posterior with the missing variable $y_{i,2}$ in

Variables	N_1	N_2	N_3	N_4
$y_{i,1}$	✓	○	✓	○
$y_{i,2}$	✓	✓	⊗	○
$y^*_{i,3}$	×	×	×	×
$y^*_{i,4}$	×	×	×	×

Table 1.2: Minimal data augmentation scheme. The symbols ✓, ×, ⊗, and ○ denote whether the variable is observed, latent but augmented, missing but augmented, or missing but not augmented in the posterior, respectively.

N_3 and the latent variables $\{y^*_{i,3}\, y^*_{i,4}\}$ for all observations, while leaving $y_{i,1}$ in $N_2 \cup N_4$ and $y_{i,2}$ in N_4 out of the sampler, as illustrated in Table 1.2. While the choices of variables and observations for augmentation appear arbitrary, they are specifically chosen to facilitate the sampling of Ω (see Section 1.3.3 for more details). By assuming that $y_{i,1}$ is missing more than $y_{i,2}$, this algorithm includes less than 50% of all missing data, which results in lower storage costs. In the vehicle choice application with 2, 297 observations, only 18% of the total missing data is used.

1.3.2 Posterior Analysis

The data-augmented posterior density is proportional to the product of the data-augmented likelihood and the prior density for the unknown parameters:

$$\pi(\theta, y_{miss}, y^*|y_{obs}) \propto f(y_{obs}, y_{miss}, y^*|\theta)\pi(\theta). \tag{1.8}$$

Define the vector $\theta = (\tilde{\beta}, \delta, \Omega)$, where $\tilde{\beta} = (\beta'_1, \beta'_2, \beta'_3, \beta'_4)'$ and $\delta = \{\delta_{t_j,j}\}$, to contain all the unknown parameters. Also, define y_{miss} and y^* to respectively contain the augmented missing outcomes and latent variables from Table 1.2, and define y_{obs} to contain all the observed data from Table 1.1.

Due to the intricate pattern of missing outcomes, specific quantities for each case of observability need to be defined. Let

$$\tilde{y}_{i,1:4} = (y_{i,1}, y_{i,2}, y^*_{i,3}, y^*_{i,4})', \tilde{y}_{i,2:4} = (y_{i,2}, y^*_{i,3}, y^*_{i,4})',$$

$$\tilde{y}_{i,134} = (y_{i,1}, y^*_{i,3}, y^*_{i,4})', \tilde{y}_{i,3:4} = (y^*_{i,3}, y^*_{i,4})',$$

and using similar notation, let $\tilde{X}_{i,1:4}$, $\tilde{X}_{i,2:4}$, $\tilde{X}_{i,134}$, and $\tilde{X}_{i,3:4}$ be block-diagonal matrices with the corresponding vectors of covariates on the block diagonals and zeros elsewhere. Similarly, define $S'_{2:4}$, S'_{134}, and $S'_{3:4}$ to be conformable matrices that "select out" the appropriate regression coefficients when pre-multiplied to $\tilde{\beta}$. For example,

$$\tilde{X}_{i,3:4} = \begin{pmatrix} x'_{i,3} & 0 \\ 0 & x'_{i,4} \end{pmatrix}, \quad S_{3:4} = \begin{pmatrix} 0 \\ I \end{pmatrix}, \quad \text{and} \quad S'_{3:4}\tilde{\beta} = \begin{pmatrix} \beta_3 \\ \beta_4 \end{pmatrix}.$$

Now, define and partition Ω and Ω_{22} as

$$\Omega = \begin{pmatrix} \underset{(1\times1)}{\Omega_{11}} & \Omega_{12} \\ \Omega_{21} & \underset{(3\times3)}{\Omega_{22}} \end{pmatrix}, \quad \Omega_{22} = \begin{pmatrix} \underset{(1\times1)}{\overline{\Omega}_{11}} & \overline{\Omega}_{12} \\ \overline{\Omega}_{21} & \underset{(2\times2)}{\overline{\Omega}_{22}} \end{pmatrix},$$

and denote the covariance matrix for $\tilde{y}_{i,134}$ as Ω_{134}.

The data-augmented likelihood needed in equation (1.8) is given by

$$\begin{aligned} f(y_{obs}, y_{miss}, y^*|\theta) \quad \propto \quad & \prod_{N_1 \cup N_3} \phi(\tilde{y}_{i,1:4}|\tilde{X}_{i,1:4}\tilde{\beta}, \Omega) \prod_{N_2} \phi(\tilde{y}_{i,2:4}|\tilde{X}_{i,2:4}S'_{2:4}\tilde{\beta}, \Omega_{22}) \times \quad (1.9) \\ & \prod_{N_4} \phi(\tilde{y}_{i,3:4}|\tilde{X}_{i,3:4}S'_{3:4}\tilde{\beta}, \overline{\Omega}_{22}) \prod_{i=1}^{N} \prod_{j=3}^{4} \mathbb{I}(\alpha_{y_{i,j}-1,j} < y^*_{i,j} \leq \alpha_{y_{i,j},j}), \end{aligned}$$

where $\phi(x|\mu, \Sigma)$ denotes the density of a multivariate normal distribution with mean vector μ and covariance matrix Σ, and $\mathbb{I}(\cdot)$ denotes an indicator function. The last product in (1.9) is the joint probability function of the ordered selection variables, which is known with certainty conditional on the latent variables. For some calculations, the data-augmented likelihood marginally of the missing outcomes is needed; it is obtained by integrating $\{y_{i,2}\}_{i \in N_3}$ out of equation (1.9) and is given by

$$\begin{aligned} f(y_{obs}, y^*|\theta) \quad \propto \quad & \prod_{N_1} \phi(\tilde{y}_{i,1:4}|\tilde{X}_{i,1:4}\tilde{\beta}, \Omega) \prod_{N_2} \phi(\tilde{y}_{i,2:4}|\tilde{X}_{i,2:4}S'_{2:4}\tilde{\beta}, \Omega_{22}) \times \\ & \prod_{N_3} \phi(\tilde{y}_{i,134}|\tilde{X}_{i,134}S'_{134}\tilde{\beta}, \Omega_{134}) \prod_{N_4} \phi(\tilde{y}_{i,3:4}|\tilde{X}_{i,3:4}S'_{3:4}\tilde{\beta}, \overline{\Omega}_{22}) \times \\ & \prod_{i=1}^{N} \prod_{j=3}^{4} \mathbb{I}(\alpha_{y_{i,j}-1,j} < y^*_{i,j} \leq \alpha_{y_{i,j},j}). \end{aligned} \tag{1.10}$$

Prior independence is assumed for simplicity. Let

$$\tilde{\beta} \sim \mathcal{N}(\beta_0, B_0), \quad \Omega \sim \mathcal{IW}(\nu_1, Q), \quad \delta \sim \mathcal{N}(\delta_0, D_0), \tag{1.11}$$

where the priors for $\tilde{\beta}$ and δ are multivariate normal, and the prior for Ω is inverse-Wishart. The hyperparameters are set to reflect prior information. To be non-informative, set the mean vectors β_0 and δ_0 to zeros, the covariance matrices B_0 and D_0 to diagonal matrices with 100 on the diagonals, ν_1 to 4, and Q to an identity matrix.

1.3.3 Sampling Algorithm

For the computations that will follow, define δ_j and $\delta_{(-j)}$ to contain all the transformed cutpoints for equations j and other than j, respectively. Similarly, define y_j^* and $y_{(-j)}^*$ to contain the latent variables from y^* for equations j and other than j.

The posterior distribution is simulated by MCMC methods. The algorithm, which omits extraneous quantities from the conditioning set, is summarized as follows:

1. Sample $\tilde{\beta}$ from the distribution $\tilde{\beta}|y_{obs}, \Omega, y^*$.
2. Sample (δ_j, y_j^*) for $j = 3, 4$ from the distribution $\delta_j, y_j^*|y_{obs}, \tilde{\beta}, \Omega, \delta_{(-j)}, y_{(-j)}^*$.
3. Sample Ω from the distribution $\Omega|y_{obs}, \tilde{\beta}, y_{miss}, y^*$.
4. Sample $y_{i,2}$ for $i \in N_3$ from the distribution $y_{i,2}|y_{obs}, \tilde{\beta}, \Omega, y^*$.

Note that the quantities $\tilde{\beta}$, δ_j, and y_j^* are sampled without conditioning on the missing outcomes as this improves the mixing of the Markov chain. As the number of iterations approaches infinity, the draws can be shown to come from the posterior distribution of interest by collapsed MCMC theory (Liu, 1994) and Metropolis-Hastings convergence results (Chib and Greenberg, 1995; Tierney, 1994).

Identification in the ordered probit equations is achieved by imposing multiple cutpoint restrictions, following Fang (2008) and Jeliazkov et al. (2008). The cutpoints $\alpha_{1,j}$ and $\alpha_{2,j}$ are fixed at zero and one, respectively, along with $\alpha_{0,j} = -\infty$ and $\alpha_{T_j,j} = +\infty$. The proposed restrictions offer two advantages. First, the elements of Ω corresponding to the ordered variables are not restricted to be in correlation form, which allows for straightforward interpretation. Second, the transformed cutpoints do not need to be sampled when the selection variables only have three categories.

Sampling $\tilde{\beta}$

The conditional distribution for $\tilde{\beta}$ can be easily derived by combining (1.10) and the normal prior for $\tilde{\beta}$. By completing the square in the exponential functions, the distribution of interest can be recognized as $\mathcal{N}(\overline{\beta}, \overline{B})$, where

$$\overline{\beta} = \overline{B}\left(\begin{array}{c} \sum_{N_1} \tilde{X}'_{i,1:4}\Omega^{-1}\tilde{y}_{i,1:4} + \sum_{N_2} S_{2:4}\tilde{X}'_{i,2:4}\Omega_{22}^{-1}\tilde{y}_{i,2:4} + \\ \sum_{N_3} S_{134}\tilde{X}'_{i,134}\Omega_{134}^{-1}\tilde{y}_{i,134} + \sum_{N_4} S_{3:4}\tilde{X}'_{i,3:4}\overline{\Omega}_{22}^{-1}\tilde{y}_{i,3:4} + B_0^{-1}\beta_0 \end{array} \right),$$

$$\overline{B} = \left(\begin{array}{c} \sum_{N_1} \tilde{X}'_{i,1:4}\Omega^{-1}\tilde{X}_{i,1:4} + \sum_{N_2} S_{2:4}\tilde{X}'_{i,2:4}\Omega_{22}^{-1}\tilde{X}_{i,2:4}S'_{2:4} + \\ \sum_{N_3} S_{134}\tilde{X}'_{i,134}\Omega_{134}^{-1}\tilde{X}_{i,134}S'_{134} + \sum_{N_4} S_{3:4}\tilde{X}'_{i,3:4}\overline{\Omega}_{22}^{-1}\tilde{X}_{i,3:4}S'_{3:4} + B_0^{-1} \end{array} \right)^{-1}.$$

Sampling (δ_j, y_j^*)

The pair (δ_j, y_j^*) is sampled in one block from the joint distribution

$$\delta_j, y_j^* | y_{obs}, \tilde{\beta}, \Omega, \delta_{(-j)}, y^*_{(-j)}$$

for $j = 3, 4$, as proposed in Albert and Chib (2001) and Chen and Dey (2000). The vector of transformed cutpoints δ_j is first sampled marginally of y_j^* from

$$\delta_j | y_{obs}, \tilde{\beta}, \Omega, \delta_{(-j)}, y^*_{(-j)},$$

and then y_j^* is sampled conditionally on δ_j from

$$y_j^* | y_{obs}, \tilde{\beta}, \Omega, \delta, y^*_{(-j)}.$$

Sampling is performed jointly, because drawing δ_j and y^*_j each from their full conditional distributions may induce high autocorrelation in the MCMC chains (Nandram and Chen, 1996).

The marginal distribution of δ_j, recovered by integrating y^*_j out of the joint distribution, is difficult to sample from directly. Instead, an independence chain Metropolis-Hastings step is used. A new draw, δ'_j, is proposed from a multivariate t distribution with $\nu_2 = 5$ degrees of freedom, $f_T(\delta_j|\hat{\delta}_j, \hat{D}_j, \nu_2)$, where $\hat{\delta}_j$ and $\hat{D}_j$ are the maximizer and negative Hessian of $f(y_j|y_{\text{obs}(-j)}, \tilde{\beta}, \Omega, \delta_j, y^*_{(-j)})\pi(\delta_j|\delta_{(-j)})$ evaluated at the maximum, respectively. The vectors y_j and $y_{\text{obs}(-j)}$ respectively contain all elements in y_{obs} associated with equations j and other than j. The acceptance probability for δ'_j is

$$\alpha_{MH}(\delta_j, \delta'_j) = min\left\{1, \frac{f(y_j|y_{\text{obs}(-j)}, \tilde{\beta}, \Omega, \delta'_j, y^*_{(-j)})\pi(\delta'_j|\delta_{(-j)})f_T(\delta_j|\hat{\delta}_j, \hat{D}_j, \nu_2)}{f(y_j|y_{\text{obs}(-j)}, \tilde{\beta}, \Omega, \delta_j, y^*_{(-j)})\pi(\delta_j|\delta_{(-j)})f_T(\delta'_j|\hat{\delta}_j, \hat{D}_j, \nu_2)}\right\}, \quad (1.12)$$

where the conditional probabilities of y_j can be calculated as products of univariate normal distribution functions (Chib et al., 2009, Section 2.1).

By independence across observational units, the vector y^*_j can be recovered by sampling $y^*_{i,j}$ from $y^*_{i,j}|y_{obs}, \tilde{\beta}, \Omega, \delta, y^*_{(-j)}$ for $i = 1, \ldots, N$. From equation (1.10), this distribution is truncated normal. Let $\mathcal{TN}(\mu, \sigma^2, a, b)$ denote a univariate normal distribution truncated to the region (a, b) with mean μ and variance σ^2. The distribution of interest is given by

$$y^*_{i,j}|y_{obs}, \tilde{\beta}, \Omega, \delta, y^*_{(-j)} \sim \mathcal{TN}(\mu_{i,j}, \sigma^2_{i,j}, \alpha_{y_{i,j-1},j}, \alpha_{y_{i,j},j}), \quad (1.13)$$

where $\mu_{i,j}$ and $\sigma^2_{i,j}$ are the conditional mean and variance for a normal distribution.

Sampling Ω

Due to the non-standard form of the posterior density in equation (1.8), the covariance matrix Ω cannot be sampled in one block from the usual inverse-Wishart distribution. Instead, one-to-one transformations of Ω and Ω_{22} will be sampled and used to construct a draw for Ω. The presented derivations are an extension of Chib et al. (2009) by applying two sets of transformations instead of one due to the additional incidentally truncated outcome.

Define the transformations

$$\Omega_{11\cdot2} = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}, \quad B_{21} = \Omega_{22}^{-1}\Omega_{21},$$

$$\overline{\Omega}_{11\cdot2} = \overline{\Omega}_{11} - \overline{\Omega}_{12}\overline{\Omega}_{22}^{-1}\overline{\Omega}_{21}, \quad \overline{B}_{21} = \overline{\Omega}_{22}^{-1}\overline{\Omega}_{21},$$

and partition Q and Q_{22} as

$$Q = \begin{pmatrix} \underset{(1\times1)}{Q_{11}} & Q_{12} \\ Q_{21} & \underset{(3\times3)}{Q_{22}} \end{pmatrix}, Q_{22} = \begin{pmatrix} \underset{(1\times1)}{\overline{Q}_{11}} & \overline{Q}_{12} \\ \overline{Q}_{21} & \underset{(2\times2)}{\overline{Q}_{22}} \end{pmatrix}.$$

To sample Ω_{22}, a change of variables from Ω_{22} to $(\overline{\Omega}_{22}, \overline{\Omega}_{11\cdot2}, \overline{B}_{21})$ is applied to the density $\Omega_{22}|y_{obs}, \tilde{\beta}, y^*$ with Jacobian $|\overline{\Omega}_{22}|$. The resulting density is proportional to a product of three recognizable distribution kernels, namely two inverse-Wisharts and one matric-normal. They are

$$\begin{aligned}
\overline{\Omega}_{22}|y_{obs}, \tilde{\beta}, y^* &\sim \mathcal{IW}(\nu_1 + N - 1, \overline{Q}_{22} + \sum_{i=1}^{N} \tilde{\epsilon}_{i,3:4}\tilde{\epsilon}'_{i,3:4}), && (1.14)\\
\overline{\Omega}_{11\cdot 2}|y_{obs}, \tilde{\beta}, y^* &\sim \mathcal{IW}(\nu_1 + n_1 + n_2, \overline{R}_{11\cdot 2}), && (1.15)\\
\overline{B}_{21}|\overline{\Omega}_{11\cdot 2}, y_{obs}, \tilde{\beta}, y^* &\sim \mathcal{MN}_{(2\times 1)}(\overline{R}_{22}^{-1}\overline{R}_{21}, \overline{\Omega}_{11\cdot 2} \otimes \overline{R}_{22}^{-1}), && (1.16)
\end{aligned}$$

where

$$\tilde{\epsilon}_{i,3:4} = (\tilde{y}_{i,3:4} - \tilde{X}_{i,3:4}S'_{3:4}\tilde{\beta}),$$

$$\tilde{\epsilon}_{i,2:4} = (\tilde{y}_{i,2:4} - \tilde{X}_{i,2:4}S'_{2:4}\tilde{\beta}),$$

$$R_{22} = (Q_{22} + \sum_{N_1 \cup N_2} \tilde{\epsilon}_{i,2:4}\tilde{\epsilon}'_{i,2:4})$$

is partitioned to be conformable with Q_{22} using similar notation, and

$$\overline{R}_{11\cdot 2} = \overline{R}_{11} - \overline{R}_{12}\overline{R}_{22}^{-1}\overline{R}_{21}.$$

By drawing from (1.14) to (1.16) and manipulating the inverted quantities, a draw of Ω_{22} marginally of the missing data can be recovered.

To sample Ω, a similar change of variables from Ω to $(\Omega_{22}, \Omega_{11\cdot 2}, B_{21})$ is applied to $\Omega|y_{obs}, \tilde{\beta}, y_{miss}, y^*$ with a Jacobian of $|\Omega_{22}|$. The resulting distributions of interest are

$$\begin{aligned}
\Omega_{11\cdot 2}|y_{obs}, \tilde{\beta}, y_{miss}, y^* &\sim \mathcal{IW}(\nu_1 + n_1 + n_3, R_{11\cdot 2}), && (1.17)\\
B_{21}|\Omega_{11\cdot 2}, y_{obs}, \tilde{\beta}, y_{miss}, y^* &\sim \mathcal{MN}_{(3\times 1)}(R_{22}^{-1}R_{21}, \Omega_{11\cdot 2} \otimes R_{22}^{-1}), && (1.18)
\end{aligned}$$

where $\tilde{\epsilon}_{i,1:4} = (\tilde{y}_{i,1:4} - \tilde{X}_{i,1:4}\tilde{\beta})$, $R = (Q + \sum\limits_{N_1 \cup N_3} \tilde{\epsilon}_{i,1:4}\tilde{\epsilon}'_{i,1:4})$ is partitioned to be conformable with Q, and $R_{11\cdot 2} = R_{11} - R_{12}R_{22}^{-1}R_{21}$. The covariance matrix Ω can now be recovered using draws from (1.14) to (1.18).

Sampling $\tilde{y}_{i,2}$

From (1.8), the conditional distributions of $\tilde{y}_{i,2}$ are easily recognized as

$$y_{i,2}|y_{obs}, \tilde{\beta}, \Omega, y^* \sim \mathcal{N}(\eta_i, \omega_i^2) \text{ for } i \in N_3, \tag{1.19}$$

where η_i and ω_i^2 are the conditional mean and variance of $y_{i,2}$.

1.4 Simulation Study

This section evaluates the performance of the MDA algorithm from Section 1.3.3 using simulated data. For reference, the algorithm is compared to a similar one that augments all the missing data from Table 1.1, denoted as y^*_{miss}, and conditions on them at every step. Specifically, this benchmark algorithm with full data augmentation (FDA) recursively samples from $[\tilde{\beta}|y_{obs}, y^*_{miss}, \Omega, y^*]$, $[\delta_j, y^*_j|y_{obs}, y^*_{miss}, \tilde{\beta}, \Omega, \delta_{(-j)}, y^*_{(-j)}]$, $[\Omega|y_{obs}, y^*_{miss}, \tilde{\beta}, y^*]$, and $[y^*_{miss}|y_{obs}, \tilde{\beta}, \Omega, \delta, y^*]$, which are densities similar to those found in Section 1.3.3, but without the missing data integrated out.

The model being considered is from equations (1.1) to (1.6). It contains five explanatory variables for each equation (including the constant), three ordered categories for each selection variable, and $N = 500$ observations. The unknown parameters are chosen so that the pattern of missing data is similar to the application in Section

1.5, with roughly 50% and 25% of the outcomes are missing from equations (1.1) and (1.2), respectively. As a result, the MDA algorithm only augments about 20% of the total missing data. The covariance matrix Ω is set to $I + 0.7ii'$, where i is a column of ones, implying a correlation of about 0.4 between the equations. Let ω_{ij} denote the element of Ω corresponding to the i-th row and j-th column. The prior hyperparameters are chosen to be non-informative as discussed in Section 1.3.2. Note that various combinations of observations, explanatory variables, and proportions of missing data were considered, but the results are not presented as they did not vary much from the performance pattern in this base case.

The algorithms are iterated 12,500 times with a burn-in of 2,500 draws. The MDA algorithm takes 17 seconds to perform 1,000 iterations on average, and the FDA algorithm takes 19 seconds. Following Chib et al. (2009), the performance is studied with inefficiency factors over 30 Monte Carlo replications. The inefficiency factor for the k-th parameter is defined as $1 + 2\sum_{l=1}^{L} \rho_k(l)(1 - \frac{l}{L})$, where $\rho_k(l)$ is the sample autocorrelation at the l-th lag for the k-th parameter, and L is the lag in which the autocorrelations taper off (Chib, 2001). This quantity measures the efficiency loss when using correlated MCMC samples instead of independent samples; values close to 1 generally indicate an efficient sampler.

Before the results are discussed, note that the sampling density for $\overline{\Omega}_{22}$ does not depend on y_{miss}, while the densities involving the other elements of Ω do depend on the missing data through $\tilde{y}_{i,1:4}$ and $\tilde{y}_{i,2:4}$ in R and R_{22}, respectively. This suggests that the benefits of MDA relative to FDA will be more evident for the elements of Ω involving equations (1.1) and (1.2) since they depend on the missing data. Using a similar argument, the elements of β_1 and β_2 from $\tilde{\beta}$ are also expected to have larger gains in performance relative to β_3 and β_4.

Boxplots of the inefficiency factors from both algorithms are displayed in Figure 1.1.

The plots for $\tilde{\beta}$ suggest that these parameters are generally sampled efficiently, except for the constant in β_1 due to the large fraction of missing outcomes in equation (1.1). The median inefficiency factors for β_3 and β_4 (excluding the constant) are between 1.1 and 1.4 under both algorithms. As previously discussed, the two algorithms perform similarly here, although the inefficiency factors are slightly lower for MDA. For β_1 and β_2, the median inefficiency factors are around 1 under MDA, but fall between 1.5 and 3 under FDA, indicating that MDA provides very efficient draws. Even with the inclusion of more explanatory variables (e.g. 21 variables for each equation, as in the empirical application), the resulting inefficiency factors under MDA are close to 1 (the results are not included due to graphical limitations). For Ω, the inefficiency factors are predictably lower when using MDA. For example, the median factors for ω_{11} and ω_{31} decrease from 9 to 3 and 17 to 4, respectively.

To show that the sampling of Ω is correct under MDA, the posterior means for $vech(\Omega)$ from one of the Monte Carlo replication are

$$(1.67, 0.76, 1.74, 0.74, 0.67, 1.69, 0.71, 0.71, 0.65, 1.64)'.$$

The corresponding posterior standard deviations are

$$(0.09, 0.06, 0.06, 0.08, 0.05, 0.08, 0.06, 0.05, 0.04, 0.06)'.$$

As the number of observations increase, the posterior means approach the true Ω from the data generating process.

Overall, the median and average inefficiency factors for all the parameters estimated using the MDA algorithm are less than or equal to their FDA counterparts. This result is consistent with the notion that, in this context, data augmentation is only used to increase the tractability of the sampling densities, so integrating them out of

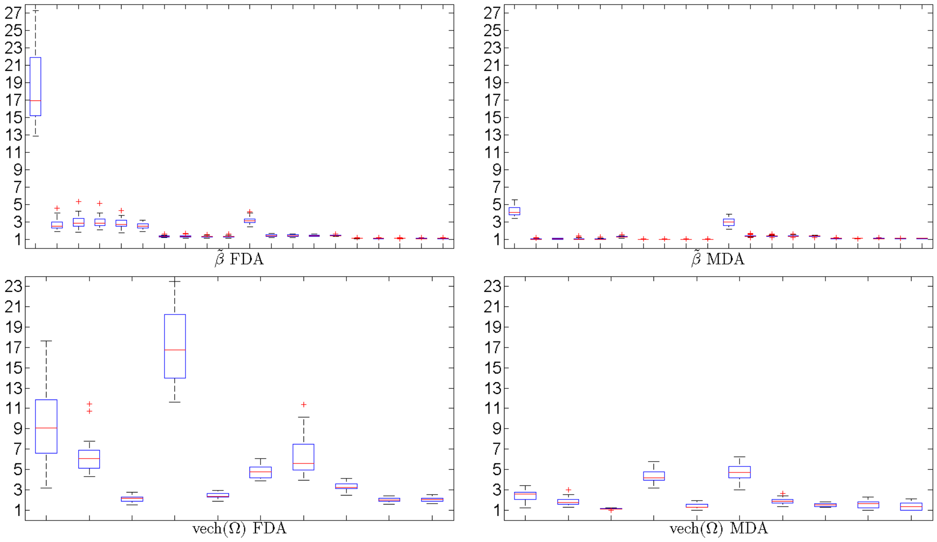

Figure 1.1: Boxplots of inefficiency factors using the FDA and MDA algorithms for $\tilde{\beta} = (\beta_1', \beta_2', \beta_3', \beta_4')'$ and $vech(\Omega) = (\omega_{11}, \omega_{21}, \omega_{22}, \omega_{31}, \omega_{32}, \omega_{33}, \omega_{41}, \omega_{42}, \omega_{43}, \omega_{44})'$.

the densities should not reduce the performance of the algorithm. For the majority of the parameters, the MDA algorithm results in lower inefficiency factors, indicative of both lower autocorrelations between the MCMC draws and of improved sampler performance when only a minimal subset of missing data is included. This result is especially evident for the parameters that are highly dependent on the missing data. As for the remaining parameters, they are efficiently estimated in both algorithms.

1.5 Application

Studies suggest that higher urban spatial structure, including residential density, is related to lower vehicle usage (Brownstone and Fang, 2009; Brownstone and Golob, 2009; Cervero and Kockelman, 1997; Dunphy and Fisher, 1996; Fang, 2008). As a result, residential density is one parameter in reducing fuel consumption of automobiles or influencing household travel behavior. Policies targeting residential density can complement traditional ones such as limiting vehicle usage by total mileage driven or enforcing fuel efficiency on vehicles. Improved understanding of this relationship can also influence city development, zoning decisions, congestion growth, and project evaluations. However, vehicle usage data commonly contains a large proportion of missing values due to the lack of vehicle ownership. If these missing values are not modeled correctly or simply omitted from the sample, estimates of interest will suffer from misspecification errors.

The sample selection model is used to jointly study the effects of residential density on vehicle usage and holdings in California. One possible causal relationship suggests that denser areas increase the cost of operating vehicles. Residential areas with more houses per square mile commonly have narrow streets, congested roads, and limited parking spaces, which contribute to higher vehicle fuel consumption and operating costs when traveling around these neighborhoods, especially for less fuel-efficient vehicles. As a result, households will tend to drive less on average and switch to more fuel-efficient vehicles. The data is obtained from the 2001 National Household Travel Survey from which a subsample 2,297 households from California is used. Table 1.3 provides detailed summary statistics. Outcomes of interest are the annual mileage driven with trucks and cars (measures of vehicle usage) and the number of trucks and cars owned by a household (measures of vehicle holdings). They are modeled jointly with exogenous covariates such as residential density, household size, income, home ownership status, and education levels.

Variable	Description	Mean	SD
	Dependent variables		
$TMILE$	Mileage per year driven with trucks (1,000 miles)	7.14	10.97
$CMILE$	Mileage per year driven with cars (1,000 miles)	8.90	10.00
$TNUM$	Number of trucks owned by the household	0.72	0.79
$CNUM$	Number of cars owned by the household	1.10	0.82
	Exogenous covariates		
$DENSITY$	Houses per square mile	2564.99	1886.09
$BIKES$	Number of bicycles	0.97	1.23
$HHSIZE$	Number of individuals in a household	2.69	1.44
$ADLTS$	Number of adults in a household	1.99	0.79
URB	Household is in an urban area	0.93	0.25
$INC1$	Household income is between 20K and 30K	0.11	0.31
$INC2$	Household income is between 30K and 50K	0.21	0.41
$INC3$	Household income is between 50K and 75K	0.19	0.39
$INC4$	Household income is between 75K and 100K	0.13	0.33
$INC5$	Household income is greater than 100K	0.22	0.41
$HOME$	Household owns the home	0.69	0.46
HS	Highest household education is a high school degree	0.31	0.46
BS	Highest household education is at least a bachelor's degree	0.46	0.50
$CHILD1$	Youngest child is under 6 years old	0.17	0.37
$CHILD2$	Youngest child is between 6 and 15 years old	0.18	0.38
$CHILD3$	Youngest child is between 15 and 21 years old	0.06	0.23
LA	Household lives in Los Angeles MSA	0.42	0.49
SAC	Household lives in Sacramento MSA	0.08	0.27
SD	Household lives in San Diego MSA	0.09	0.28
SF	Household lives in San Francisco MSA	0.23	0.42

Table 1.3: Descriptive statistics based on $2,297$ observations.

The model is given by

$$\begin{aligned}
y_{i,1} &= \beta_{0,1} + log(DENSITY_i)\beta_{1,1} + x_i'\beta_1 + \epsilon_{i,1}, \qquad (1.20)\\
y_{i,2} &= \beta_{0,2} + log(DENSITY_i)\beta_{1,2} + x_i'\beta_2 + \epsilon_{i,2},\\
y^*_{i,3} &= \beta_{0,3} + log(DENSITY_i)\beta_{1,3} + x_i'\beta_3 + \epsilon_{i,3},\\
y^*_{i,4} &= \beta_{0,4} + log(DENSITY_i)\beta_{1,4} + x_i'\beta_4 + \epsilon_{i,4},
\end{aligned}$$

for $i = 1, \ldots, 2,297$ households, where $y_{i,1}$ and $y_{i,2}$ are annual mileage driven with trucks and cars, $y^*_{i,3}$ and $y^*_{i,4}$ are the latent variable representations for the number of trucks and cars owned ($y_{i,3}$ and $y_{i,4}$), and x'_i is a row vector of exogenous covariates from Table 1.3. The equation subscript j is omitted from x'_i since the same covariates are used in every equation, and the covariate $log(DENSITY_i)$ is separated to emphasize that it is a variable of interest. The error structure is $(\epsilon_{i,1}, \epsilon_{i,2}, \epsilon_{i,3}, \epsilon_{i,4})' \sim \mathcal{N}(0, \Omega)$. The selection variables are the number of trucks and cars a household owns, which have categories of zero, one, or more than two. Sample selection is modeled as follows: $y_{i,1}$ is observed if and only if $y_{i,3} > 0$, and $y_{i,2}$ is observed if and only if $y_{i,4} > 0$. Grouping households that own more than two trucks and cars (2.26% and 4.48% of the sample, respectively) with households that own two trucks and cars is for estimation convenience, because the transformed cutpoints do not need to be sampled. The two combined groups are assumed to be similar, so this grouping should not affect the analysis.

The model estimates are in Table 1.4, and the marginal effects with respect to residential density are in Table 1.5. The quantities of interest are obtained by iterating the algorithm 110,000 times, discarding the first 10,000 iterations for burn-in, and taking the ergodic averages over the associated draws. Prior hyperparameters are set to reflect non-informativeness since the effects of residential density and other covariates are not known a priori.

For the truck and car mileage equations, the posterior means for the coefficients of $log(DENSITY)$ are -0.41 and -0.25 with posterior standard deviations of 0.32 and 0.23, respectively. The signs suggest that households located in denser neighborhoods, all else equal, are associated with lower truck and car usage on average. For example, the marginal effects from Table 1.5 show that a 50% increase in residential density is associated with a 168.18 and 98 decrease in annual mileage driven with trucks and

Variable	$TMILE$		$CMILE$		$TNUM$		$CNUM$	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$log(DENSITY)$	-0.41	(0.32)	-0.25	(0.23)	-0.07	(0.02)	-0.02	(0.02)
$BIKES$	-0.16	(0.28)	0.03	(0.20)	0.08	(0.02)	-0.01	(0.01)
$HHSIZE$	0.45	(0.52)	0.73	(0.42)	0.05	(0.03)	-0.06	(0.03)
$ADLTS$	-0.63	(0.68)	0.28	(0.53)	0.09	(0.04)	0.17	(0.03)
URB	0.43	(1.48)	-0.69	(1.22)	-0.14	(0.08)	0.19	(0.08)
$INC1$	2.53	(1.67)	-1.35	(1.02)	0.18	(0.08)	0.09	(0.06)
$INC2$	1.28	(1.46)	1.15	(0.88)	0.41	(0.07)	0.11	(0.05)
$INC3$	2.56	(1.49)	1.65	(0.91)	0.49	(0.07)	0.26	(0.06)
$INC4$	2.60	(1.60)	0.74	(1.01)	0.59	(0.08)	0.24	(0.07)
$INC5$	3.63	(1.58)	1.86	(0.97)	0.61	(0.08)	0.31	(0.06)
$HOME$	-0.61	(0.90)	-1.26	(0.56)	0.21	(0.04)	0.10	(0.04)
HS	-0.41	(0.98)	1.28	(0.70)	0.02	(0.05)	0.11	(0.04)
BS	-2.04	(1.03)	0.85	(0.71)	-0.20	(0.05)	0.17	(0.05)
$CHILD1$	1.71	(1.45)	0.56	(1.07)	0.12	(0.08)	0.12	(0.07)
$CHILD2$	1.24	(1.31)	0.61	(0.98)	0.08	(0.07)	0.06	(0.06)
$CHILD3$	1.32	(1.51)	0.01	(1.07)	0.04	(0.08)	-0.02	(0.07)
LA	2.71	(0.99)	1.51	(0.74)	-0.14	(0.05)	0.03	(0.05)
SAC	2.09	(1.40)	1.74	(1.03)	-0.15	(0.08)	0.07	(0.07)
SD	1.26	(1.42)	0.07	(1.02)	-0.18	(0.08)	0.10	(0.07)
SF	1.58	(1.17)	-0.06	(0.81)	-0.27	(0.06)	0.15	(0.05)

Table 1.4: Model estimates. Posterior means and standard deviations of the coefficients are reported.

cars, respectively. These estimates are small despite increasing residential density by as much as 50%. The results suggest that residential density has a small economic impact on vehicle usage. Also, the differences in magnitudes suggest that less fuel-efficient vehicles are more sensitive to residential density changes than fuel-efficient vehicles on average. The results are consistent with the intuition that households would want to drive less as overall vehicle operating costs increased, which is particularly true for less efficient vehicles. However, the posterior standard deviations are close in magnitude to the coefficient estimates, which suggest some uncertainty in the relationship between residential density and vehicle usage for trucks and cars. This finding is somewhat contrary to the conclusions in Brownstone and Fang (2009) and Fang (2008), where the vehicle usage variables are modeled as censored (Tobit-type) outcomes instead of potential outcomes. The authors find that residential density does affect truck utilization in a significant way but not for car utilization. This difference arises due to the different modeling strategies.

Marginal effects are presented in Table 1.5 since the coefficients in the ordered equations are difficult to interpret. The estimates suggest that when residential density increases by 50%, the probability of not holding any trucks increases by 1.318%, while the probability of holding one and two or more trucks decrease by 0.637% and 0.681%, respectively. The effects on car holdings is practically on the same order of magnitude, but there is sizable uncertainty in the estimates as the posterior standard deviations are large. These estimates are similar to the findings in Fang (2008) and approximately half the size of the estimates in Brownstone and Fang (2009).

$\Delta Pr(TNUM = 0)$	$\Delta Pr(TNUM = 1)$	$\Delta Pr(TNUM \geq 2)$
13.18	-6.37	-6.81
(3.35)	(1.67)	(1.72)
$\Delta Pr(CNUM = 0)$	$\Delta Pr(CNUM = 1)$	$\Delta Pr(CNUM \geq 2)$
2.88	-0.23	-2.64
(2.89)	(0.26)	(2.65)
$\Delta TMILE$	$\Delta CMILE$	
-168.14	-98.00	
(130.70)	(93.85)	

Table 1.5: Marginal effects of increasing $DENSITY$ by 50%. The changes in probabilities are in 10^{-3} units, and the changes in truck and car mileage are in annual miles.

1.6 Concluding remarks

This paper develops an efficient algorithm to estimate sample selection models with two incidentally truncated outcomes and two separate selection variables. While such models are easily described mathematically, estimation is often difficult due to the intricate pattern of missing data that arises with two incidentally truncated outcomes and the discrete nature of the selection variables. These features result in evaluations of intractable likelihoods, identification issues, and computationally-inefficient algorithms. This paper extends the Markov chain Monte Carlo algorithm with minimal data augmentation, first proposed in Chib (2007) and Chib et al. (2009), to efficiently simulate the joint posterior distribution of interest. A central aspect of the proposed algorithm is that it includes only a small subset of the total missing data in the MCMC sampler, resulting in improved sampling efficiency and decreased computational load, as demonstrated in the simulation study. Also, despite not having the "complete" data, the resulting sampling distributions are well-known and easy to sample from.

The model is applied to estimate the effects of residential density on vehicle usage and holdings in the state of California. Results suggest that large increases in residential density are not strongly associated with changes in either vehicle utilization or the probability of holding cars, but are strongly related to changes in truck holdings. This finding associated with vehicle utilization, especially for truck usage, is contrary to the literature and demonstrates that the sample selection framework can reveal new conclusions in the data.

Chapter 2

Bayesian analysis of multivariate sample selection models using Gaussian copulas

2.1 Introduction

This paper applies Bayesian methods to estimate a multivariate sample selection model that addresses the ubiquitous problem of sample selection. In general, sample selection occurs when a variable of interest is non-randomly missing for a subset of the sample, resulting in a sample that is not representative of the desired population. A well-known application involves analyzing market wages that are only partially observed, depending on whether the individual is participating in the labor force or not. If inference is based only on the remaining observed sample, then specification errors may arise.

A widely used model to address sample selection involves modeling an observed binary selection variable, y_1, that determines whether a continuous outcome variable, y_2, is missing or observed (Heckman, 1976, 1979a). Because the joint distribution for (y_1, y_2) is difficult to specify, the model is often re-parameterized in terms of y_1^* and y_2, where y_1^* is a continuous and latent representation of y_1, with the distributional assumption that $(y_1^*, y_2) \sim \mathcal{N}_2(\mu, \Sigma)$. The joint normality assumption is made to achieve tractable results and to obtain an explicit measure of dependence between the two variables through Σ.

Although many variations of this model have been developed and estimated for selection and outcome variables with different data types (e.g. count, ordered, censored, etc.) and distributional assumptions, they are often limited to the specific distributions assumed in the corresponding papers or to formulations with only two or three variables. For example, Terza (1998) studies a univariate count data regression subject to a binary selection variable, and Boyes et al. (1989) analyzes a binary regression with a separate binary selection variable. From a Bayesian perspective, Chib et al. (2009), Greenberg (2007), and van Hasselt (2009) provide analyses for a single Tobit or binary selection variable. For extensive surveys on other variations of sample selection models from a non-Bayesian perspective, refer to Vella (1998) and Greene (2008). However, certain theoretical and applied problems may require either different distributional assumptions or more dependent variables than these models and methods can accommodate, which limits the problems that can be studied.

To address these issues, we analyze a flexible multivariate sample selection model in which the desired marginal distributions are specified by the practitioner. The multivariate dependence is modeled through a copula function in conjunction with the specified marginal distributions. A copula, broadly speaking, is a function that links a multivariate distribution function to its univariate distribution functions with a particular dependence structure (Sklar, 1959). In other words, there exists a copula function $\mathbb{C}$ such that $F(y_1, \ldots, y_n) = \mathbb{C}(F_1(y_1), \ldots, F_n(y_n))$, where $F(y_1, \ldots, y_n)$ is a multivariate distribution function with $F_1(y_1), \ldots, F_n(y_n)$ as the univariate distribution functions. This method is particularly useful when $F_1(y_1), \ldots, F_n(y_n)$ are known and $F(y_1, \ldots, y_n)$ is unknown, because the copula provides an alternative representation of the joint density.

This paper uses Gaussian copulas that are constructed using multivariate normal distribution functions and a theorem from Sklar (1959). While copulas have been used extensively in the statistics literature for several decades, their usage in econometrics has been relatively limited. Early work on copulas include Hoeffding (1940), Fréchet (1951), and Sklar (1959, 1973), with the latter proving an important theorem that states all continuous multivariate distribution functions have a unique copula representation; the reader is referred to Nelsen (1998) and Zimmer and Trivedi (2005) for comprehensive treatments on copula theory. For multivariate Gaussian copulas, Lee (2010) and Pitt et al. (2006) respectively analyze multivariate count data models and general regression models using Bayesian simulation methods.

Recent work from econometrics pertaining to sample selection and copulas include Bhat and Eluru (2009), Genius and Strazzera (2008), Lee (1983), Prieger (2000), Smith (2003), and Zimmer and Trivedi (2006). Lee (1983) does not impose joint normality on the standard sample selection model but uses a bivariate Gaussian copula to link the two specified marginal distributions together. Similarly, Prieger (2000) and Bhat and Eluru (2009) develop a model based on a Farlie-Gumbel-Morgenstern copula, which only has moderate correlation coverage between the selection and outcome variables. The remaining authors analyze selection models using variations of Archimedean copulas, resulting in closed-form expressions that are relatively simple to estimate. The aforementioned papers on selection models mostly use maximum likelihood estimation and stay within a bivariate or trivariate structure.

This paper has two purposes. First, we analyze and estimate a multivariate sample selection model with p pairs of selection and outcome variables using Gaussian copulas, where each variable may be discrete or continuous with any parametric marginal distribution specified by the practitioner. We thereby move beyond the bivariate or trivariate structure of the preceding papers to accommodate a larger class of sample selection models. Second, we show how the Bayesian Markov chain Monte Carlo (MCMC) simulation methods from Lee (2010) and Pitt et al. (2006) can be applied to accommodate copula models with missing data. The proposed estimation method has two main advantages. By using Bayesian simulation methods, it is not necessary to repeatedly compute the high-dimensional copula distribution functions that are needed with non-Bayesian methods. Even though there are methods to calculate these distribution functions (Bőrsch-Supan and Hajivassiliou, 1993; Geweke, 1991; Hajivassiliou and McFadden, 1998; Jeliazkov and Lee, 2010; Keane, 1994), the resulting likelihood is difficult to maximize, even for low-dimensional problems (Zimmer and Trivedi, 2005). Next, our proposed algorithm does not require simulation of the missing data and their associated quantities, which has been shown to improve the efficiency of the Markov chain (Chib et al., 2009; Li, 2011). Careful consideration is needed in this context since the amount and complexity of missing data grow simultaneously with the number of variables modeled (e.g. a model with p partially observed outcomes can have 2^p different combinations of missing data for each observation).

The methods are applied to study the effects of residential density on vehicle miles traveled and vehicle holdings for households in California. Residential density and household demographic variables are used to explain the number of miles a household drives with trucks and cars and the number of trucks and cars a household owns.

The rest of the paper is organized as follows. Section 2.2 provides a brief introduction to copulas, and Section 2.3 describes the proposed multivariate sample selection model. Section 2.4 presents the estimation algorithm while Section 2.5 illustrates the methods on simulated and actual data. The paper is concluded in Section 2.6.

2.2 Copulas

This section provides a brief introduction to copulas. Intuitively, a copula is a function that links a multivariate joint distribution to its univariate distribution functions. This approach allows joint modeling of outcomes for which the multivariate distributions are difficult to specify, which is often the case in econometric modeling (e.g. models for discrete choice, count data, and combinations of discrete and continuous data).

More formally, a copula $\mathbb{C}$ has the following definition from Zimmer and Trivedi (2005):

Definition 1 *An n-dimensional copula (or n-copula) is a function $\mathbb{C}$ from the unit n-cube $[0,1]^n$ to the unit interval $[0,1]$ which satisfies the following conditions:*

1. $\mathbb{C}(1,\ldots,1,u_k,1,\ldots,1) = u_k$ *for every $k \leq n$ and for all u_k in $[0,1]$;*
2. $\mathbb{C}(u_1,\ldots,u_n) = 0$ *if $u_k = 0$ for any $k \leq n$;*
3. $\mathbb{C}$ *is n-increasing*

From this definition, a copula can be viewed as an n-dimensional distribution function for $U_1,\ldots,U_n$ defined over $[0,1]^n$, where U_i is uniformly distributed over $[0,1]$ $(i = 1,\ldots,n)$.

An important result is that multivariate distribution functions can be expressed in terms of a copula function and its univariate distribution functions. Let $Y_1,\ldots,Y_n$

be n continuous random variables with an n-dimensional distribution function

$$F(y_1, \ldots, y_n)$$

and marginal distribution functions $F_1(y_1), \ldots, F_n(y_n)$. Then,

$$\begin{aligned} F(y_1, \ldots, y_n) &= Pr(Y_1 < y_1, \ldots, Y_n < y_n) \qquad (2.1) \\ &= Pr(U_1 < F_1(y_1), \ldots, U_n < F_n(y_n)) \qquad (2.2) \\ &= \mathbb{C}(u_1 = F_1(y_1), \ldots, u_n = F_n(y_n)) \qquad (2.3) \end{aligned}$$

since $F_i(Y_i) \sim U_i$ by the integral transformation result. The dependence between the marginal distributions is introduced through a dependence parameter specific to the chosen copula function. Note that the copula function in (2.3) is unique if $F_1(y_1), \ldots, F_n(y_n)$ are continuous distribution functions. The relationship in (2.3) still holds for discrete distributions, but the copula function is not unique.

Although many copulas exist, this paper uses a multivariate Gaussian copula of the form

$$\mathbb{C}(u_1, \ldots, u_n | \, \Omega \,) = \Phi_n(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n) | \, \Omega \,), \tag{2.4}$$

where $\Phi_n(\cdot)$ is an n-dimensional distribution function for a multivariate normal vector z with mean zero and correlation matrix Ω, and $\Phi^{-1}(\cdot)$ is the inverse distribution function of a univariate standard normal random variable. Intuitively, the proposed approach is to transform the original variables with pre-specified margins into uniform random variables and then into a new set of correlated random variables, z, that is distributed $\mathcal{N}(0, \Omega)$. The advantage of this approach is that dependence is easier to handle through the transformed data z than through the original or uniform random variables. From Song (2000), the density for this copula is proportional to

$$|\Omega|^{-\frac{1}{2}} \exp(0.5\, z'(I - \Omega^{-1})z), \tag{2.5}$$

where $z_i = \Phi^{-1}(u_i)$ $(i = 1, \ldots, n)$, and I is an identity matrix with the same dimensions as Ω.

The Gaussian copula has several desirable properties. It is one of the few multivariate copulas with $\frac{n(n-1)}{2}$ dependence parameters (the off-diagonals of Ω), one for each pair of variables. This feature is especially desirable in this context since the dependence between the selection and outcome variables is of interest. Furthermore, unlike some copulas, the dependence measures for this copula can be positive or negative. This property is also attractive as the sign of the dependence between the selection and outcome variables is not known a priori. Lastly, Gaussian copulas attain the Fréchet lower and upper bounds when the dependence parameters approach -1 and 1, respectively. This last property is an important factor when choosing a copula and implies that the Gaussian copula can cover the space between the Fréchet bounds.

2.3 Model

Suppose we have $2p$ variables with N observations for each. Let the first p variables denote the selection variables that determine whether the remaining p variables of interest are observed. Following Pitt et al. (2006), for observational units $i = 1, \ldots, N$

Variables	Case 1	Case 2	Case 3	Case 4
$y_{i,1}$	✓	✓	✓	✓
$y_{i,2}$	✓	✓	✓	✓
$y_{i,3}$	✓	○	✓	○
$y_{i,4}$	✓	✓	○	○

Table 2.1: Four cases of variable observability when $p = 2$. The symbols ○ and ✓ respectively denote whether the variable is missing or observed.

and variables $j = 1, \ldots, 2p$, the proposed model is

$$y_{i,j} = F_{i,j}^{-1}(\Phi(z_{i,j})), \quad z_i = (z_{i,1}, \ldots, z_{i,2p})' \overset{\text{iid}}{\sim} \mathcal{N}_{2p}(0, \Omega). \tag{2.6}$$

In equation (2.6), $y_{i,j}$ is a discrete or continuous variable with distribution function $F_{i,j}(\cdot)$ that depends on the covariates $x_{i,j}$ and a vector of unknown parameters $\theta_{i,j}$. Also, denote $f_{i,j}(\cdot|\theta_{i,j})$ as the density function for $y_{i,j}$. As an example, if $y_{i,j} \sim \mathcal{N}(x_{i,j}'\beta, \sigma^2)$, then $\theta_{i,j} = (\beta, \sigma^2)$, and $F_{i,j}(\cdot)$ is a distribution function for a normal random variable with mean $x_{i,j}'\beta$ and variance σ^2. Furthermore, each $y_{i,j}$ is modeled with a corresponding Gaussian latent variable $z_{i,j}$, along with a column vector z_i that is distributed multivariate normal with mean zero and correlation matrix Ω. For simplicity, let $F_{i,j}(\cdot) = F_j(\cdot)$, $f_{i,j}(\cdot|\theta_{i,j}) = f_j(\cdot|\theta_j)$, and $\theta_{i,j} = \theta_j$ for all i and j, which imply that the j-th distribution function and vector of unknown parameters are the same across all observational units. For (2.6), it is important to note that the mapping $F_{i,j}^{-1}(\Phi(\cdot))$ is one-to-one when $y_{i,j}$ is continuous and many-to-one when $y_{i,j}$ is discrete. Also, for identification reasons, the set of covariates for each selection variable should include at least one additional covariate than the corresponding variable of interest.

Sample selection is incorporated by assuming that the first p selection variables, $y_{i,1}, \ldots, y_{i,p}$, are always observed and determine whether the remaining variables of interest, $y_{i,p+1}, \ldots, y_{i,2p}$, are missing or observed. That is, $y_{i,1}$ determines whether $y_{i,p+1}$ is missing or observed, $y_{i,2}$ determines whether $y_{i,p+2}$ is missing or observed, etc. This paired sample selection structure implies that for any observational unit i, there are 2^p possible combinations of missing variables. Table 2.1 illustrates the combinations when $p = 2$. We observe either $(y_{i,1}, y_{i,2}, y_{i,3}, y_{i,4})$, $(y_{i,1}, y_{i,2}, y_{i,4})$, $(y_{i,1}, y_{i,2}, y_{i,3})$, or $(y_{i,1}, y_{i,2})$. In the context of the transportation economics application, $y_{i,1}$ and $y_{i,2}$ are the number of trucks and cars the i-th household owns, and $y_{i,3}$ and $y_{i,4}$ are the mileage driven with these vehicles. The mileage variables are missing when the number of vehicles is zero.

2.4 Estimation

2.4.1 Posterior density and priors

The data-augmented posterior density of interest is proportional to the data-augmented likelihood multiplied by the prior densities: $\pi(\theta, \Omega, z|\, y) \propto f(z, y|\theta, \Omega)\pi(\theta, \Omega)$. In this expression, θ contains θ_j for all variables, y contains all the observed $y_{i,j}$ variables, and z contains all the Gaussian latent variables from the copula function corresponding to y. The form of $f(z, y|\theta, \Omega)$ will be described in the next subsection. For Bayesians, this posterior density summarizes all the information available for the unknown parameters after seeing the data. It combines prior information on the parameters before seeing data with information from the observed data through the likelihood function.

We assume independent priors such that $\pi(\theta, \Omega) = \pi(\theta)\, \pi(\Omega)$ for convenience. The prior for Ω is $\mathcal{IW}(\nu, Q)$, an inverse-Wishart distribution with scalar hyperparameter ν and $2p \times 2p$ hyperparameter Q. Because θ is application-specific, we will leave prior specifications to the practitioner. Note that conjugate priors do not aid tractability when using copulas in this context. Therefore, we suggest practitioners choose priors with simple functional forms that accurately reflect prior knowledge and proper priors if model comparisons with Bayes factors are desired.

2.4.2 Estimation algorithm

The posterior distribution is approximated by MCMC methods, largely following Lee (2010) and Pitt et al. (2006). For the algorithm that follows, define y_j and z_j to be the elements of y and z corresponding to the j-th variable, respectively. Also, let z_{-j} be $z\backslash z_j$ and θ_{-j} be $\theta\backslash\theta_j$. The algorithm to sample from $\pi(\theta, \Omega, z|\, y)$ is summarized as follows:

1. Sample Ω in one block from $f(\Omega|z)$.
2. Sample (θ_j, z_j) jointly for all discrete marginal distributions from

   $$f(\theta_j, z_j|y, z_{-j}, \theta_{-j}, \Omega)$$

   as follows

   (a) Sample θ_j without z_j from $f(\theta_j|y, z_{-j}, \theta_{-j}, \Omega)$.

   (b) Sample z_j conditioned on θ_j from $f(z_j|y, z_{-j}, \theta, \Omega)$.
3. Sample θ_j for all continuous marginal distributions from $f(\theta_j|y, z_{-j}, \theta_{-j}, \Omega)$ and solve for z_j with y_j and θ_j through the one-to-one transformation in (2.6).

The Metropolis-Hastings algorithm is used to sample from the preceding distributions since random draws cannot be easily obtained from the posterior distribution using direct sampling. Broadly speaking, this algorithm generates samples from the posterior distribution by first proposing candidate values from a known proposal distribution and then accepting them with a certain probability. If a proposed value is rejected, then the previous value is used. This method constructs a Markov chain such that after a sufficient burn-in period, the draws can be shown to come from the posterior distribution of interest by Metropolis-Hastings convergence results (Chib and Greenberg, 1995; Tierney, 1994) as the number of iterations approaches infinity.

A multivariate t proposal distribution with mean μ, scale matrix V, and degrees of freedom ν is used in the subsequent sections. In order to obtain parameter values such that this proposal density dominates the density of interest (also called the target density), μ and V are set to the maximum and inverse of the negative Hessian (evaluated at the maximum) of the density of interest, respectively. These quantities can be obtained by quasi-Newton methods. Lastly, the degrees of freedom parameter ν is set to ensure heavy tails. Specific details are provided in the subsequent sampling sections.

Note that the missing variables of interest (e.g. the entries marked with a ○ in Table 2.1) and their corresponding Gaussian latent variables are not sampled. In many Bayesian MCMC algorithms for missing data problems, the missing data are often included in the sampling to facilitate the tractability of the sampling densities, however this strategy is not necessary and not always optimal. Chib et al. (2009) and Li (2011) have shown that the inclusion and conditioning of missing data in some applications can slow down the mixing of the Markov chain. In particular, for the semiparametric sample selection model of Chib et al. (2009), the inefficiency factors (a measure of how quickly the autocorrelations in a Markov chain chain taper off, where lower values indicate better performance) are at least 4 times greater when the missing data due to sample selection are included in the sampler. This issue is particularly problematic when the quantity of missing outcomes due to sample selection is large or when the model includes many parameters, both of which may be the case in this context. Therefore, the proposed algorithm does not sample the missing data and corresponding latent data.

Sampling Ω

Two different algorithms are presented to sample Ω. The first algorithm is based on the sampler from Chib and Greenberg (1998); it works well when the number of variables is small (less than 4) and is easy to implement. For problems with more than 4 variables, we introduce an algorithm based on Chan and Jeliazkov (2009).

Since the observed variables can potentially change for every observational unit, additional notation will now be defined. Let s_i denote the indices of the observed variables for observation i, and let y_{s_i} and z_{s_i} respectively denote the columns of observed and latent variables corresponding to s_i such that $Var(z_{s_i}) = \Omega_{s_i}$. For example, if $y_{i,1}$, $y_{i,2}$, and $y_{i,4}$ are observed for $p = 2$, then $s_i = \{1, 2, 4\}$, $y_{s_i} = (y_{i,1}, y_{i,2}, y_{i,4})'$, $z_{s_i} = (z_{i,1}, z_{i,2}, z_{i,4})'$, and

$$\Omega_{s_i} = \begin{pmatrix} 1 & \omega_{12} & \omega_{14} \\ \omega_{21} & 1 & \omega_{24} \\ \omega_{41} & \omega_{42} & 1 \end{pmatrix}.$$

The full conditional density $f(\Omega|z)$ is proportional to

$$f(z|\Omega)\pi(\Omega) \propto \pi(\Omega) \prod_{i=1}^{N} \left\{ |\Omega_{s_i}|^{-\frac{1}{2}} \exp(-0.5\, z'_{s_i} \Omega_{s_i}^{-1} z_{s_i}) \right\}. \tag{2.7}$$

Since Ω is a $2p \times 2p$ correlation matrix, there are $\frac{2p(2p-1)}{2}$ unique off-diagonal terms, denoted by ω, that need to be sampled. To sample ω from (2.7), a Metropolis-Hastings step with a multivariate t proposal is used. The target density in (2.7) is first maximized with respect to ω using quasi-Newton methods; let $\widehat{\omega}$ and $\widehat{V}$ denote the maximizing vector and the inverse of the negative Hessian evaluated at the maximum. Next, propose ω' from a multivariate t distribution with mean vector $\widehat{\omega}$, scale matrix $\widehat{V}$, and degrees of freedom ν. A proposed value for Ω' can now be constructed with ω'. If Ω' is not positive definite, then the previous value of Ω is used instead. Otherwise, the draw is accepted with probability

$$\alpha(\omega, \omega') = \min\left\{1, \frac{f(\Omega'|z) f_T(\omega|\widehat{\omega}, \widehat{V}, \nu)}{f(\Omega|z) f_T(\omega'|\widehat{\omega}, \widehat{V}, \nu)}\right\}. \tag{2.8}$$

The second algorithm is based on the sampling strategy from Chan and Jeliazkov (2009). To introduce the technique, note that any positive definite covariance matrix Σ can be decomposed as $\Sigma = L' D^{-1} L$. The unit lower triangular matrix L contains ones on the diagonal and unrestricted elements on the lower off-diagonal, while the diagonal matrix D contains positive elements on the diagonal and zeros elsewhere. The insight of this algorithm is that we can sample the elements in L and D instead of the elements in Σ directly and reconstruct Σ through the decomposition.

Using similar notation to Chan and Jeliazkov (2009), denote λ_j $(j = 1, \ldots, 2p)$ as the diagonal elements of D and $a_{j,k}$ $(1 \leq k < j \leq 2p)$ as the unrestricted elements on the lower off-diagonal of L. Similarly, denote $a^{j,k}$ as the (j,k)-th element of L^{-1}. As an illustration,

$$D = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_{2p} \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ a_{2,1} & 1 & 0 & \ldots & 0 \\ a_{3,1} & a_{3,2} & 1 & \ldots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{2p,1} & a_{2p,2} & \ldots & \ldots & 1 \end{pmatrix}.$$

Let $\sigma_{j,k}$ denote the element of Σ corresponding to the j-th row and k-th column. After imposing $\sigma_{1,1} = \sigma_{2,2} = \ldots = \sigma_{2p,2p} = 1$ in Σ to obtain correlation form and expanding $L'D^{-1}L$, the free elements of Σ must satisfy the constraints

$$\sigma_{j,k} = a^{j,k}\lambda_k + \sum_{h=1}^{k-1} a^{j,h}a^{k,h}\lambda_h, \quad 1 \leq k < j \leq 2p, \tag{2.9}$$

and λ_j must satisfy

$$\lambda_1 = 1, \tag{2.10}$$

$$\lambda_j = 1 - \sum_{k=1}^{j-1}(a^{j,k})^2\lambda_k, \quad j = 2, \ldots, 2p. \tag{2.11}$$

As noted in the referenced paper, when Σ is in correlation form, $\{\lambda_j\}$ is only a function of $\{a_{j,k}\}$. This implies that the off-diagonal elements in $\{\sigma_{j,k}\}$ are also functions of $\{a_{j,k}\}$ only. Consequently, we only need to sample $\{a_{j,k}\}$ when Σ is expressed as $L'D^{-1}L$.

The second algorithm will also utilize a Metropolis-Hastings step since (2.7) is not a recognizable distribution with respect to $\{a_{j,k}\}$. First, decompose Ω as $L'D^{-1}L$

and express (2.7) in terms of $\{a_{j,k}\}$ analogously to (2.9) through (2.11). Let $\widehat{a}$ and $\widehat{A}$ respectively denote the values that maximize (2.7) with respect to $a = \{a_{j,k} : 1 \leq k < j \leq 2p\}$ and the inverse of the negative Hessian evaluated at the maximum. Next, propose a' from a multivariate t distribution with mean $\widehat{a}$, scale matrix $\widehat{A}$, and degrees of freedom ν, which can be used to construct the proposed value Ω'. The proposed value is accepted with probability

$$\alpha(a, a') = \min \left\{ 1, \frac{f(\Omega'|z) f_T(a|\widehat{a}, \widehat{A}, \nu)}{f(\Omega|z) f_T(a'|\widehat{a}, \widehat{A}, \nu)} \right\}. \tag{2.12}$$

This approach differs slightly from the one in Chan and Jeliazkov (2009) since (2.7) is of a different form due to the missing data.

Both of these algorithms allow Ω to be sampled in one block with the positive definite constraint intact. The first algorithm is relatively easy to implement, however it is generally inefficient when the dimension of Ω is large. The positive definite constraint will become increasingly difficult to satisfy when p increases, resulting in proposed values that are frequently rejected and slower mixing of the Markov chain. The second algorithm is more involved but has been shown to be efficient (Chan and Jeliazkov, 2009), therefore it is recommended for models with more than 4 variables.

Sampling (θ_j, z_j) for discrete marginals

For discrete marginals, the pair (θ_j, z_j) is sampled jointly. Using the method of composition, θ_j is first sampled from $f(\theta_j|y, z_{-j}, \theta_{-j}, \Omega)$, then z_j is sampled from

$f(z_j|y, z_{-j}, \theta, \Omega)$ with θ_j conditioned on. The first density is

$$f(\theta_j|\, y, z_{-j}, \theta_{-j}, \Omega) \;\propto\; \pi(\theta_j|\theta_{-j}) \prod_{i=1}^{N} \{f(y_{i,j}|\, z_{-j}, \theta_j, \Omega)\}^{\mathbb{I}(j \in s_i)}, \tag{2.13}$$

where

$$f(y_{i,j}|\, z_{-j}, \theta_j, \Omega) \;=\; \int f(y_{i,j}|z_{i,j}, \theta_j) f(z_{i,j}|z_{-j}, \Omega)\, dz_{i,j}, \tag{2.14}$$

and $\mathbb{I}(A)$ is an indicator function that takes the value 1 when A is true and 0 otherwise. Upon defining $\mu_{i,j|-j}$ and $\sigma^2_{i,j|-j}$ as the conditional mean and variance of the normal density $f(z_{i,j}|z_{-j}, \Omega)$, it can be shown that

$$f(y_{i,j}|\, z_{-j}, \theta_j, \Omega) = \Phi\left(\frac{T^U - \mu_{i,j|-j}}{\sigma_{i,j|-j}}\right) - \Phi\left(\frac{T^L - \mu_{i,j|-j}}{\sigma_{i,j|-j}}\right), \tag{2.15}$$

with

$$\begin{aligned} T^L &= \Phi^{-1}(F_j(y_{i,j} - 1)), & (2.16) \\ T^U &= \Phi^{-1}(F_j(y_{i,j})). & (2.17) \end{aligned}$$

The sampling of θ_j can proceed with the target density in (2.13) and a Metropolis-Hastings step just like the preceding sections.

Now, the density for $z_{i,j}$ is

$$f(z_{i,j}|y, z_{-j}, \theta, \Omega) \propto f(y_{i,j}|z_{i,j}, \theta_j) f(z_{i,j}|z_{-j}, \Omega), \tag{2.18}$$

where

$$f(y_{i,j}|z_{i,j}, \theta_j) \quad = \quad \mathbb{I}(T^L < z_{i,j} \leq T^U).$$

Thus, for any $j \in S_i$ that corresponds to a discrete $y_{i,j}$,

$$z_{i,j}|y, z_{-j}, \theta, \Omega \sim \mathcal{TN}_{(T^L, T^U)}(\mu_{i,j|-j}, \sigma^2_{i,j|-j}), \tag{2.19}$$

where $\mathcal{TN}_{(a,b)}(\mu, \sigma^2)$ denotes a univariate normal distribution with mean μ and variance σ^2 truncated to the region (a, b). Note that the conditional moments depend on which latent variables from z_{-j} are available for observation i, indicated by s_i, and need to be adjusted accordingly.

Sampling θ_j for continuous marginals

For continuous marginal distributions, θ_j is sampled from $f(\theta_j|y, z_{-j}, \theta_{-j}, \Omega)$. The sampling density is proportional to

$$\pi(\theta_j|\theta_{-j}) \prod_{i=1}^{N} \{f(y_{i,j}|\theta_j)\}^{\mathbb{I}(j \in S_i)} \exp(0.5\, z'_{S_i}(I_{S_i} - \Omega^{-1}_{S_i}) z_{S_i}). \tag{2.20}$$

Note that the elements in z_{S_i} corresponding to the j-th variable are also functions of θ_j, so the last term cannot be dropped. This is from the relationship $z_{i,j} = \Phi^{-1}(F_j(y_{i,j}))$, where $F_j(y_{i,j})$ depends on θ_j. A Metropolis-Hastings step is needed to obtain a draw from (2.20). Once θ_j is drawn, the elements in z_j can be recovered through the aforementioned relationship, therefore z_j does not need to be sampled for continuous marginals.

2.5 Applications

2.5.1 Simulated data

This section illustrates the estimation methods with simulated data. The purpose is to study the performance of the algorithm on a model that will be used in the next subsection and to demonstrate that the algorithm can correctly recover the parameters of interest. Specifically, the model from Section 2.3 is estimated with two Poisson selection variables ($y_{i,1}$ and $y_{i,2}$) and two normally-distributed outcome variables ($y_{i,3}$ and $y_{i,4}$). To set the context, sample selection is incorporated as follows: $y_{i,3}$ is observed if and only if $y_{i,1} > 0$, and $y_{i,4}$ is observed if and only if $y_{i,2} > 0$.

For $i = 1, \ldots, 1000$, we have the following

$$\begin{array}{lcl} y_{i,1} & \sim & Po(\lambda_{i,1}), \quad log(\lambda_{i,1}) = x'_{i,1}\beta_1, \\ y_{i,2} & \sim & Po(\lambda_{i,2}), \quad log(\lambda_{i,2}) = x'_{i,2}\beta_2, \\ y_{i,3} & \sim & \mathcal{N}(x'_{i,3}\beta_3, \sigma_3^2), \\ y_{i,4} & \sim & \mathcal{N}(x'_{i,4}\beta_4, \sigma_4^2), \end{array} \tag{2.21}$$

where $x'_{i,j}$ $(j = 1, \ldots, 4)$ are randomly-drawn exogenous covariate vectors from standard normal distributions. The true generating values for the parameters of interest $\theta_1 = \beta_1$, $\theta_2 = \beta_2$, $\theta_3 = (\beta_3, \sigma_3^2)$, $\theta_4 = (\beta_4, \sigma_4^2)$, and Ω are presented in Table 2.2. The percentage of missing data for each outcome variable is 20%, similar to the real data. Proper priors are used with hyperparameters that reflect non-informativeness. For θ_1 and θ_2, multivariate normal priors are used with mean vector zero and a variance-covariance matrix of an identity matrix multiplied by 100; similar priors are used for β_3 and β_4. Lastly, inverse gamma priors are used for σ_3^2 and σ_4^2.

The algorithm is iterated 5,000 times with 500 iterations discarded for burn-in. Table 2.2 reports the posterior means and standard deviations along with their true generated values, and Figure 2.1 illustrates the lagged autocorrelations for a randomly-chosen parameter $\beta_{1,3}$ up to order 40. In general, the results from Table 2.2 suggest that all the parameters have been estimated well since the posterior means are reasonably close to their generated values with tight standard deviations. Furthermore, the autocorrelation plot, a way of assessing how well the Markov chain mixes, suggests that our algorithm performs well. The autocorrelations for $\beta_{1,3}$ decrease and taper off around lag 40, as do most of the autocorrelations for the remaining parameters. However, we suggest iterating the algorithm at least 15,000 times to obtain more precise results.

2.5.2 Vehicle usage in California

The model from (2.21) is applied to analyze the effects of residential density on household vehicle usage in California. A sample selection framework is utilized since vehicle usage is non-randomly missing from the sample data with a probability that depends on whether the household owns a vehicle or not. Households may be selecting

Parameter	Generated value	$\mathbb{E}(\cdot\|y)$	$Std(\cdot\|y)$
$\beta_{1,1}$	0.30	0.31	0.06
$\beta_{1,2}$	0.30	0.28	0.05
$\beta_{1,3}$	0.30	0.30	0.04
$\beta_{1,4}$	0.30	0.31	0.05
$\beta_{1,5}$	0.30	0.30	0.04
$\beta_{2,1}$	0.20	0.20	0.06
$\beta_{2,2}$	0.20	0.17	0.05
$\beta_{2,3}$	0.20	0.23	0.05
$\beta_{2,4}$	0.20	0.19	0.05
$\beta_{2,5}$	0.20	0.21	0.05
$\beta_{3,1}$	0.50	0.81	0.23
$\beta_{3,2}$	0.50	0.54	0.07
$\beta_{3,3}$	0.50	0.40	0.06
$\beta_{3,4}$	0.50	0.44	0.07
σ_3^2	3.00	2.79	0.16
$\beta_{4,1}$	0.30	0.30	0.16
$\beta_{4,2}$	0.30	0.35	0.05
$\beta_{4,3}$	0.30	0.38	0.05
σ_4^2	2.00	1.92	0.10
$\omega_{2,1}$	0.28	0.25	0.04
$\omega_{3,1}$	0.28	0.27	0.03
$\omega_{3,2}$	0.28	0.27	0.04
$\omega_{4,1}$	0.28	0.31	0.04
$\omega_{4,2}$	0.28	0.28	0.04
$\omega_{4,3}$	0.28	0.27	0.04

Table 2.2: Posterior means and standard deviations for θ_j ($j = 1, \ldots, 4$) and $vech(\Omega) = (\omega_{2,1}, \omega_{3,1}, \omega_{3,2}, \omega_{4,1}, \omega_{4,2}, \omega_{4,3})$.

themselves into being vehicle owners for unobserved reasons that also affect how much they drive, creating differences in the observed and unobserved samples. Therefore, sample selection must be accounted for.

Some studies suggest that certain changes in urban spatial structure (e.g. residential density) may be effective in reducing fuel consumption of automobiles or in influencing travel behavior (Brownstone and Fang, 2009; Brownstone and Golob, 2009; Cervero and Kockelman, 1997; Dunphy and Fisher, 1996; Fang, 2008). For example, it may be more costly to maneuver around a location with higher residential density due

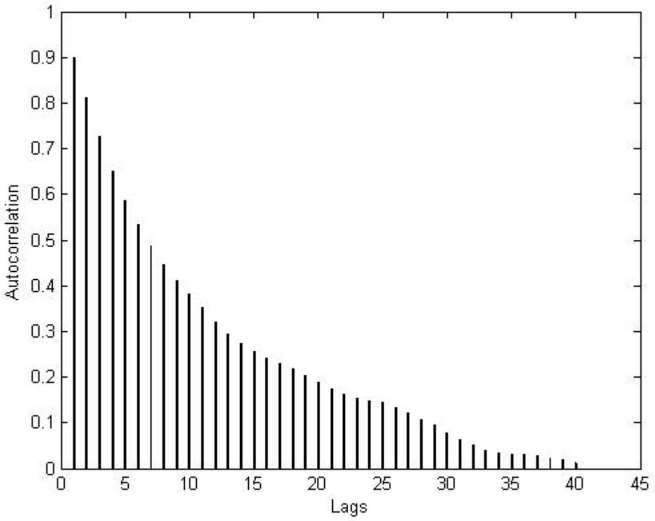

Figure 2.1: Autocorrelation plot for $\beta_{1,3}$.

to increased congestion and time spent in searching for parking spaces, resulting in households driving less and switching to more fuel efficient vehicles. Consequently, understanding this potential relationship can provide alternative policies to control fuel consumption and congestion.

The dataset is from the 2001 National Household Travel Survey. It contains the daily and long-distance travel information between April 2001 and May 2002 for approximately 66,000 households across the nation, along with variables such as residential density, household size, residential location type, income, education, and other household characteristics. The dataset used contains 1,000 randomly-sampled households that reside in California. The primary variables of interest are the number of trucks and cars owned by the household ($y_{i,1}$ and $y_{i,2}$) and the corresponding annual mileage driven with these vehicles ($y_{i,3}$ and $y_{i,4}$), where 20% to 30% of the mileage variables are missing. A truck is defined as a van, sports utility vehicle, or pickup truck, and a car is an automobile, car, or station wagon. These two categories have distinct differences in miles per gallon (MPG) requirements by the Corporate Average

Fuel Economy (CAFE) standards. Covariates of interest include residential density (housing units per square mile at the census block level), household size, and dummy variables representing whether the household resides in an urban location, is low income, has a young child, and owns their home. Descriptive statistics are summarized in Table 2.3.

Variable	Description	Mean	SD
	Dependent variables		
Tnum	Number of trucks owned by the household	0.72	0.79
Cnum	Number of cars owned by the household	1.10	0.82
Tmile	Mileage per year driven with trucks (10,000 miles)	0.71	1.10
Cmile	Mileage per year driven with cars (10,000 miles)	0.89	1.00
	Exogenous covariates		
Density	Houses per square mile	2564.99	1886.09
Hhsize	Number of individuals in a household	2.69	1.44
Urb	Household is in an urban area	0.93	0.25
Lowinc	Household income is between 20K and 30K	0.11	0.31
Child	Youngest child is under 6 years old	0.17	0.37
Home	Household owns the home	0.26	0.44

Table 2.3: Descriptive statistics based on 1,000 observations.

The results are presented in Tables 2.4 and 2.5. From Table 2.4, the estimated correlation between the truck equations is 0.37, suggesting that sample selection is not ignorable for these vehicles. This relationship is due to positive associations in unobserved factors that affect both truck ownership and utilization (e.g. a predisposition to travel more in spacious vehicles like trucks). On the other hand, the estimated correlation for the car equations is negligible and suggests that selection may not be an issue in this case.

1.00	-0.41	0.37	-0.02
-0.41	1.00	-0.17	-0.02
0.37	-0.17	1.00	0.01
-0.02	-0.02	0.01	1.00

Table 2.4: Posterior means for Ω

	Tnum		Cnum		Tmile		Cmile	
Covariates	Mean	SD	Mean	SD	Mean	SD	Mean	SD
log(Density)	-0.06	0.03	0.21	0.09	-0.01	0.26	0.05	0.17
Hhsize	0.09	0.05	0.12	0.61	0.15	0.23	0.09	0.16
Urb	-0.52	0.55	0.43	0.49	-0.13	1.13	-0.15	0.90
Lowinc	-0.23	0.58	-0.35	0.41	0.20	0.93	-0.07	0.74
Child	0.10	0.37	-0.24	0.32	0.10	0.78	-0.09	0.56
Home	0.26	0.27	0.34	0.20	·	·	·	·

Table 2.5: Posterior means and standard deviations of β_1, β_2, β_3, and β_4.

The estimates in Table 2.5 suggest that the effect of residential density on vehicle usage is uncertain. The posterior standard deviations are large relative to the means, and the 95% probability intervals for these parameters contain 0 (not shown in the table). This result is consistent with the findings of Li (2011) in which a multivariate sample selection model is also used to analyze a similar application. However, this conclusion differs from the ones presented in Fang (2008) and Brownstone and Golob (2009), where these authors generally find evidence for negative associations between truck usage and residential density. This difference arises due to the usage of a sample selection model and different distributional assumptions.

On the other hand, there is evidence that households residing in denser neighborhoods tend to own fewer trucks and more cars. This can be attributable to the increased costs of operating vehicles with lower fuel efficiency in these areas, resulting in preferences for cars with better fuel economy. Also, larger households tend to have more trucks, presumably because these vehicles can fit more passengers.

2.6 Concluding Remarks

This paper analyzes a multivariate sample selection model with p pairs of selection and outcome variables. A unique feature of this model is that the variables can be discrete or continuous with any parametric distribution, resulting in a large class of multivariate selection models that can be accommodated. For example, the model may involve any combination of variables that are continuous, binary, ordered, or censored. Although the joint distribution can be difficult to specify, a multivariate Gaussian copula function is used to link the marginal distributions together and handle the multivariate dependence. The proposed estimation approach relies on the MCMC-based techniques from Lee (2010) and Pitt et al. (2006) and adapts the preceding methods to a missing data setting. An important aspect of this algorithm is that it does not require simulation of the missing outcomes, which has been shown in some cases to improve the mixing of the Markov chain. The methods are applied to both simulated and real data, and the results show that the algorithm works well and can reveal new conclusions in the data.

A copy of the Matlab code to estimate the model in the real data section is available upon request.

Chapter 3

A model for broad choice data

3.1 Introduction

Discrete choice models are immensely popular in empirical work. Among the numerous applications, they are used to study the choices made by individuals for transportation modes, social interactions, recreational activities, and residential locations. However, these models may not be directly applicable when there are issues with data observability. As an illustration, suppose that it is of interest to model a household's vehicle choice at the make-model-trim level (e.g. the choice set contains a Honda Civic LX, Honda Civic Hybrid, Toyota Camry LE, and Toyota Camry XLE Hybrid) but only choice data up to the make-model level (e.g. either Honda Civic or Toyota Camry) are observed. In this case, standard discrete choice models cannot be used, because the chosen vehicle from the desired choice set is not observed.

A new model to accommodate the aforementioned situation is studied in this paper. Specifically, this paper analyzes a discrete choice model where the observed outcome is not the exact alternative chosen by a decision maker (e.g. the Civic Honda LX) but rather the broad group of alternatives in which the chosen alternative belongs to (e.g. the group of Honda Civics). I refer to this as the model for broad choice data, because the observed data only broadly represent the exact choices made by the decision makers.

Although this issue with data observability is acknowledged in the literature, it is often circumvented by redefining the choice set of interest so that standard methods can be applied. For example, instead of modeling the vehicle choices at the lower make-model-trim level, the choices at the higher make-model level are modeled. To accommodate this redefined setting, the observable attributes used in the model are either aggregated or averaged over the members of the broad groups prior to estimation. However, a major drawback of using the aggregate or average attributes is that they result in loss of precision for the parameter estimates when the group members are not homogenous with respect to their attributes. Since attributes within groups are commonly heterogenous, this approach is problematic. For example, the mileage per gallon, engine size, and maintenance cost attributes can significantly differ between hybrid and non-hybrid vehicles within the same make-model group. As such, the model for broad choice data is useful, because it is uses disaggregated data and analyzes a richer choice set structure.

The paper proceeds as follows. The model for broad choice data is formally stated in Section 3.2, and the likelihood-based quantities are derived in Section 3.3. Using the quantities from the preceding section, Section 3.4 discusses the identification issues associated with using the broad choice data. The details for maximum likelihood and Bayesian estimation of the parameters are discussed in Section 3.5, and Section 3.6 illustrates the various estimators on simulated data. Concluding remarks are in Section 3.7.

3.2 Model for broad choice data

The model specification is similar to that of a multinomial logit model and is based on random utility theory. Formally, the model is expressed as

$$\begin{aligned} U_{ij}^* &= \delta_j + x_{ij}'\beta + \epsilon_{ij}, \quad \epsilon_{ij} \overset{i.i.d.}{\sim} \text{Type 1 Extreme Value}, && (3.1) \\ Y_i^* &= j \quad \text{if} \quad U_{ij}^* > U_{ik}^* \, \forall \, k \in C = \{1, 2, \ldots, J\}, && (3.2) \\ Y_i &= m \quad \text{if} \quad Y_i^* \in C_m, && (3.3) \end{aligned}$$

for decision makers $i = 1, \ldots, N$, alternatives $j = 1, \ldots, J$, and groups $m = 1, 2, \ldots, M$.

The latent utility that decision maker i obtains from alternative j is given by U_{ij}^* in (3.1). It is a function of an "average" level of utility that is constant for alternative j across all decision makers, δ_j, a column vector of K exogenous and observable attributes, x_{ij}, a column vector of unknown coefficients, β, and an unobserved error term, ϵ_{ij}, that is distributed i.i.d. Type 1 Extreme Value. For identification purposes, the average utility for the first alternative is normalized to zero (i.e. $\delta_1 = 0$). Choosing the first alternative for the normalization is inessential, as long as there is at least one such normalization.

In (3.2), the random variable Y_i^* denotes the exact alternative chosen by decision maker i from the choice set C. The decision maker chooses alternative j if it provides the most utility among all the alternatives from the choice set. In (3.3), the variable Y_i represents the broad group of alternatives that Y_i^* is from. To be more specific about the values that Y_i can take, decompose each decision maker's choice set into M groups such that $C = \bigcup_{m=1}^{M} C_m$ and $\bigcap_{m=1}^{M} C_m = \emptyset$. Then, Y_i equals the value m

if the exact alternative chosen belongs to C_m.

An important aspect of this paper is that only the outcomes for Y_i (and not Y_i^*) are observed. I refer to the observed outcomes for Y_i as the broad choice data, because they only broadly represent the exact choices made by the decision makers. For the running example, the choice set is partitioned into two groups. The first group, C_1, contains the two Honda Civics, and the remaining group, C_2, contains the two Toyota Camrys. The exact vehicle chosen by the household from C is not observed. Instead, we only observe either a 1 or 2, or equivalently, whether the exact vehicle chosen is a type of Honda Civic or Toyota Camry.

For the remainder of the paper, I refer to (3.1) to (3.3) as the model for broad choice data. Also, I refer to (3.1) to (3.2) with Y_i^* observed for all decision makers as the model for exact choice data; this model is usually referred to as the multinomial logit model in the literature. It is important to emphasize that the two models are equivalent when each group contains only a single alternative. To see their equivalence, note that Y_i is equal to Y_i^* for all decision makers when $|C_m| = 1$ for all groups, thus the two models are identical in this case.

3.3 Likelihood function and associated quantities

This section discusses the likelihood function of the sample, score function, and Hessian matrix of the log-likelihood function for the model with broad choice data. The Hessian matrix is simple and provides insight into the likelihood function. It is also useful for the discussions on identification, information loss, and estimation in the subsequent sections.

Before discussing the likelihood function, some additional notation is needed. Define

$\delta = (\delta_2, \delta_3, \ldots, \delta_J)'$, and let $\theta = (\delta', \beta')'$ be the parameter vector with $G = J - 1 + K$ elements. Also, define $w_{ij} = (z_j', x_{ij}')'$, where z_j is a column vector of zeros and ones such that $w_{ij}'\theta$ is equal to the right hand side of the latent utility in (3.1). Intuitively, these vectors select out the appropriate average utility in θ for U_{ij}^*. As an illustration, with the $\delta_1 = 0$ normalization,

$$z_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad z_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad z_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \text{and} \quad z_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{3.4}$$

when $J = 4$. In general, there are J such vectors with $J - 1$ elements in each. They do not depend on i, because the composition of average utilities in (3.1) does not vary across decision makers.

The specifications from (3.1) to (3.3) imply that Y_i takes a value of m when one of the alternatives in group m provides the highest level of utility among the alternatives in C. Since we do not observe the exact choice, the probability of observing $Y_i = m$ is equal to the probability that any alternative in C_m may be the utility-maximizing alternative. Because these events are disjoint, this probability is equal to

$$\begin{aligned} \tilde{P}_{im} &= \Pr(Y_i = m), && (3.5) \\ &= \Pr(Y_i^* \in C_m), && (3.6) \\ &= \sum_{c \in C_m} P_{ic}, && (3.7) \end{aligned}$$

where each probability within the summation is of the logit probability form

$$P_{ic} = \Pr(Y_i^* = c) = \frac{\exp\left(w_{ic}'\theta\right)}{\sum_{j=1}^{J} \exp\left(w_{ij}'\theta\right)}. \tag{3.8}$$

The log-likelihood function of the sample is expressed as

$$L_B(\theta) = \sum_{i=1}^{N} \sum_{m=1}^{M} Y_{im} \log(\tilde{P}_{im}), \tag{3.9}$$

where Y_{im} equals one if the outcome for decision maker i is equal to m and zero otherwise. The subscript B denotes a quantity corresponding to the model with broad choice data.

Differentiation of (3.9) with respect to θ yields the score function

$$S_B(\theta) = \frac{\partial L_B(\theta)}{\partial \theta} = \sum_{i=1}^{N} \left(\sum_{m=1}^{M} Y_{im} \sum_{c \in C_m} w_{ic} P_{ic|C_m} - \sum_{j=1}^{J} w_{ij} P_{ij} \right), \tag{3.10}$$

where

$$P_{ic|C_m} = \frac{\exp\left(w_{ic}'\theta\right)}{\sum_{s \in C_m} \exp\left(w_{is}'\theta\right)}, \tag{3.11}$$

and the Hessian matrix of the log-likelihood function

$$H_B(\theta) = \frac{\partial L_B(\theta)}{\partial \theta \partial \theta'} = L - F, \tag{3.12}$$

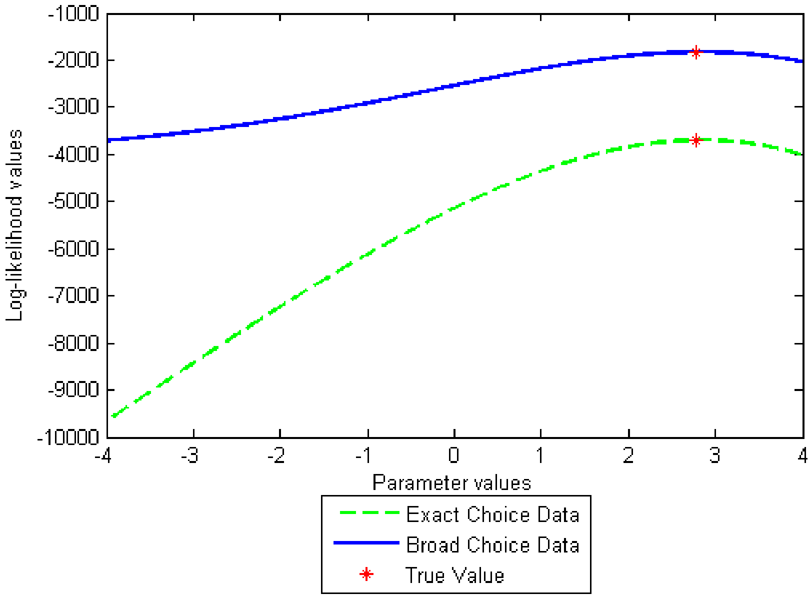

Figure 3.1: Log-likelihood functions with respect to a scalar parameter. The function corresponding to exact choice data is concave everywhere, but the one for broad choice data is not globally concave.

where

$$L = \sum_{i=1}^{N} \left(\sum_{m=1}^{M} Y_{im} \sum_{c \in C_m} (w_{ic} - \sum_{s \in C_m} P_{is|C_m} w_{is}) P_{ic|C_m} (w_{ic} - \sum_{s \in C_m} P_{is|C_m} w_{is})' \right), \tag{3.13}$$

and

$$F = \sum_{i=1}^{N} \left(\sum_{j=1}^{J} (w_{ij} - \sum_{r=1}^{J} P_{ir} w_{ir}) P_{ij} (w_{ij} - \sum_{r=1}^{J} P_{ir} w_{ir})' \right). \tag{3.14}$$

The quantity in (3.11) is interpreted as the probability of decision maker i choosing alternative $c \in C_m$ when the choice set is restricted C_m. The derivations for the preceding quantities are given in the Appendix.

The analogous quantities for the model with exact choice data, which are denoted with the E subscript, are given in McFadden (1973). The log-likelihood function of the sample is $L_E(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{J} Y_{ij}^* \log(P_{ij})$, where Y_{ij}^* equals one if j is the exact alternative chosen by decision maker i and zero otherwise. The Hessian matrix of the log-likelihood, $H_E(\theta)$, is equal to $-F$.

There are two important characteristics of $H_B(\theta)$ to emphasize. The first is that $H_B(\theta)$ is not generally negative semidefinite. To see this, note that L and F are both positive semidefinite since they are equal to weighted moment matrices of the observed attributes, but the Hessian matrix, which is equal to the difference between these two matrices, does not need to be negative semidefinite. Hence, $L_B(\theta)$ is generally not concave over the entire range of θ. On the other hand, McFadden (1973) shows that $H_E(\theta)$ is negative semidefinite, so $L_E(\theta)$ is concave in θ. Figure 3.1 confirms the shapes of the two log-likelihood functions for observed data with respect to a scalar parameter. Second, due to observing the broad choices instead of the exact choices, $L_B(\theta)$ generally has less curvature than $L_E(\theta)$. Both Figures 3.1 and 3.2 illustrate this fact. The diminished curvature has consequences on the identification of the parameters, which is discussed in the subsequent section.

3.4 Identification

Identification is assessed by analyzing whether the information matrix is nonsingular. The results in this section utilize Theorem 1 from Rothenberg (1971). Using the notation from the previous section, the theorem states that θ is locally identified if and only if the information matrix $\mathbb{I}_B(\theta) = -\mathbb{E}(H_B(\theta))$ is nonsingular, or equivalently, has rank G (the number of elements in θ).

The information matrix corresponding to (3.9) is equal to

$$\begin{aligned}\mathbb{I}_B(\theta) &= -\mathbb{E}(H_B(\theta)), && (3.15)\\ &= F - IL, && (3.16)\\ &= \mathbb{I}_E(\theta) - IL, && (3.17)\end{aligned}$$

where $F = \mathbb{I}_E(\theta)$ from McFadden (1973), and

$$IL = \sum_{i=1}^{N}\left(\sum_{m=1}^{M}\tilde{P}_{im}\sum_{c\in C_m}(w_{ic} - \sum_{s\in C_m}P_{is|C_m}w_{is})P_{ic|C_m}(w_{ic} - \sum_{s\in C_m}P_{is|C_m}w_{is})'\right). \quad (3.18)$$

This derivation is discussed in the Appendix. The expression for the information matrix in (3.17) has an intuitive form: it is equal to the difference between the information matrix with exact choice data and IL. Loosely speaking, when viewed asymptotically, IL quantifies "information loss" from using broad choice data instead of exact choice data, in the sense that $IL = \mathbb{I}_E(\theta) - \mathbb{I}_B(\theta)$ is positive definite unless $\mathbb{I}_E(\theta) = \mathbb{I}_B(\theta)$. By analyzing (3.18), the main factors that determine IL are the observed attributes, choice probabilities, and most importantly, the sizes of the groups which determine the number of outer products being summed in (3.18).

I will discuss identification of the parameters using the information matrix for three distinct cases. The first case corresponds to each group having only one alternative (i.e. $|C_m| = 1$ for all groups) which results in the model for exact choice data. It is well known that θ is identified when w_{ij} varies across alternative sets and is not collinear with the other attribute vectors (McFadden, 1973). Also, as expected, $\mathbb{I}_B(\theta)$ reduces to $\mathbb{I}_E(\theta)$ in this case. To see this, let t_m be the only element in C_m, which implies that $P_{is|C_m} = 1$ for all $s \in C_m$. Then,

$$\begin{aligned} IL &= \sum_{i=1}^{N} \left(\sum_{m=1}^{M} \tilde{P}_{im} (w_{it_m} - w_{it_m})(w_{it_m} - w_{it_m})' \right), \\ &= 0_{(G \times G)}, \end{aligned} \tag{3.19}$$

where $0_{(G\times G)}$ denotes a $G \times G$ matrix of zeros, and thus $\mathbb{I}_B(\theta) = \mathbb{I}_E(\theta)$.

The second case, presumably the most typical with broad choice data, occurs when the sizes of the groups are strictly less than J. For this case, θ is locally identified, because given the aforementioned assumptions on w_{ij}, the information matrix is full rank. It is interesting to note that IL is rank deficient since it has at most rank $G-M$, but subtracting it from $\mathbb{I}_E(\theta)$ does not decrease the rank of $\mathbb{I}_E(\theta)$. Local identification is achieved from the nonlinearities in the distributional form, log-likelihood function, and Hessian matrix with respect to the parameters and observable attributes. Global and non-parametric identification are not analyzed in this paper and are topics for future research.

Although θ is locally identified in the second case, there are scenarios in which the broad choice data do not contain much information about the parameters, leading to either imprecise estimates during estimation or a nearly singular information matrix. As alluded to earlier, the amount of information in (3.15) about θ decreases as the sizes of the groups increase. One extreme scenario in the second case occurs when there are a large number of alternatives and the size of a particular group is close to the size of C. For these group configurations, IL is "close" to $\mathbb{I}_E(\theta)$, and the information matrix is nearly singular, leading to weak identification of θ. By weak identification, I mean that although θ is locally identified, the information matrix is close to being singular. This is consistent with Figure 3.2 which depicts that, when

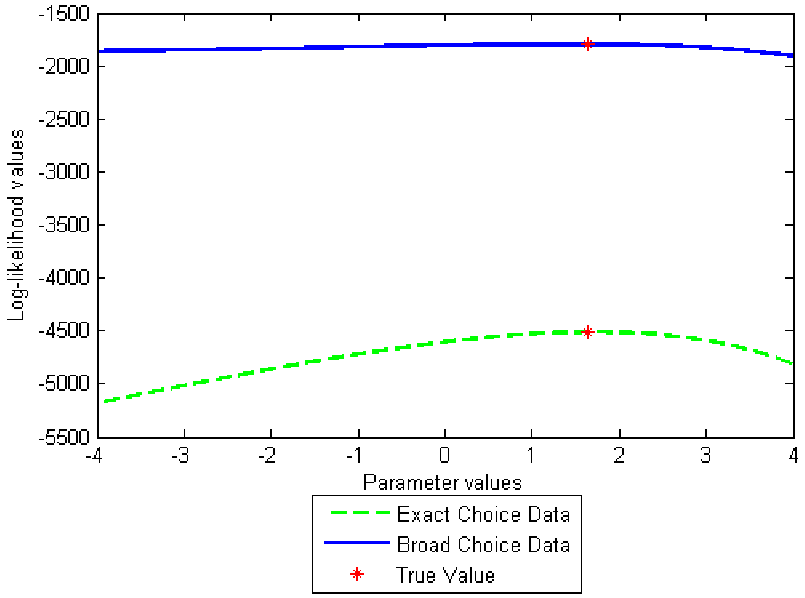

Figure 3.2: Log-likelihood functions with respect to a scalar parameter. The function corresponding to the broad choice data is almost flat.

a particular group is close in size to C, $L_B(\theta)$ is almost flat relative to $L_E(\theta)$ with respect to a scalar parameter.

The third case occurs when there is a single group that is equal to C in size, or equivalently, when every observed outcome for the broad choice data corresponds to the entire choice set. When this case occurs, θ is not identified, because the information matrix is rank deficient. To see this, assume that $m = M = 1$, which implies that $C_m = C$, $\tilde{P}_{im} = 1$ for all decision makers and groups, and $P_{is|C_m} = P_{is}$ for all $s \in C$ and decision makers. Then

$$\begin{aligned} IL &= \sum_{i=1}^{N} \left(\sum_{c \in C} (w_{ic} - \sum_{s \in C} P_{is} w_{is}) P_{ic} (w_{ic} - \sum_{s \in C} P_{is} w_{is})' \right), & (3.20) \\ &= \mathbb{I}_E(\theta), & (3.21) \end{aligned}$$

and therefore $\mathbb{I}_B(\theta) = 0_{(G\times G)}$ is rank deficient. This result establishes that only knowing a decision maker's choice belongs to the full choice set is not sufficient to discern the possible values of θ.

3.5 Estimation

This section describes maximum likelihood (ML) and Bayesian estimation of θ. Throughout the entire discussion, I assume the second case mentioned in Section 3.4 which implies that θ is at least locally identified. But for scenarios in which the broad choice data are not informative about θ (see Section 3.4 for an example), I propose incorporating additional information in the form of population market shares into the problem.

For ML estimation, the market share information is implemented as constraints on θ. To begin the discussion, assume that the population market shares for alternatives 2 through J are known. These market shares, denoted by s_j for $j = 2, 3, \ldots, J$, are defined as the percentage shares of the population choosing each alternative. They are collectively denoted by s and are related to the parameters by the nonlinear market share constraints $N^{-1}\sum_{i=1}^{N} P_{ij} = s_j$ for $j = 2, 3, \ldots, J$. The population shares are informative for the parameters, because for a well-specified model, the predicted market shares from a large representative sample should equal the population shares on average. In other words, the predicted market shares are unbiased and consistent estimates for the population market shares.

There are two important assumptions concerning these constraints. The first is that the predicted in-sample market shares, $N^{-1}\sum_{i=1}^{N} P_{ij}$, must equal the population market shares. As mentioned earlier, this assumption should be met for a well-specified model with a large representative sample on average, but equality in small samples is uncertain. The second assumption is that the population market shares are known with certainty and hence fixed. This assumption may not always hold as the population market shares may be measured with error or only known with a certain degree of certainty to the researcher. But for the purposes of ML estimation, I assume that these assumptions are not violated.

To account for possible violations in the two preceding assumptions, I propose using Bayesian methods. In contrast to the ML methods, the proposed Bayesian methods formally account for the uncertainty in the market shares, and they do not strictly enforce the constraints onto the parameters. Instead, the constraints, which contain information for θ through the market shares, are only used to construct an informative prior for θ. As a result, the prior can be used to indirectly reflect uncertainty for these constraints or equivalently for the two preceding assumptions. This is discussed in subsequent sections.

In summary, for each estimation method, I discuss how it can be implemented with or without incorporating information from the market share. Because the former case is relatively non-standard, a majority of the discussion is dedicated to it.

3.5.1 Maximum likelihood estimation

To enforce the constraints in an ML estimation routine, I use a result from Berry et al. (1995). They proved that, conditional on β and s, the constraints are one-to-one mappings that relate δ to s and β. Thus, assuming that the constraints can be inverted to solve for δ as a function of β, which I denote as $\delta(\beta)$, the maximum likelihood estimate (MLE) is obtained by maximizing the log-likelihood function for

Method Name	Form for $h(\delta^{(k)})$
BLP	$I_{(J\times J)}$
ANJ	$-[f'(\delta^{(k)})]^{-1}$
APNJ	$-[a(s)]^{-1}$
ADJ	$-Diag(f'(\delta^{(k)}))^{-1}$
APDJ	$-Diag(a(s))^{-1}$

Table 3.1: Different forms for $h(\delta^{(k)})$. In terms of convergence speed for the iterative system in (3.23), the ranking from fastest to slowest is as follows: ANJ, APNJ, ADJ, APDJ, and BLP. The specific expressions for $h(\delta^{(k)})$ in the context of (3.1) to (3.3) are given in the Appendix.

the observed sample

$$L_B(\beta, \delta(\beta)) = \sum_{i=1}^{N} \sum_{m=1}^{M} y_{im} \log(\tilde{P}_{im}) \tag{3.22}$$

with respect to β, where y_{im} is the observed value for Y_{im}, and the δ typically in $\tilde{P}_{im}$ is replaced with $\delta(\beta)$. It is tempting to interpret this approach as concentrating δ out of the likelihood, but the constraints (conditional on β and s) are not the first order conditions of (3.9) with respect to δ. As such, this approach is only used to enforce the constraints.

The main difficulty of obtaining the MLE is in inversion of the constraints. An analytic inverse is not obvious, so numerical methods must be used. I use the iterative techniques from Chapter 4 of this dissertation which are more computationally efficient than the ones presented in Berry et al. (1995) by an order of magnitude. The techniques rely on solving the market share constraints for a fixed point using an accelerated iterative system. When the system is iterated enough times, the iterates will converge to a unique vector for δ that solves the market share constraints.

The iterative system is given by

$$\delta^{(k+1)} = \delta^{(k)} + h(\delta^{(k)})f(\delta^{(k)}), \tag{3.23}$$

where the $(k+1)$ and (k) superscripts respectively indicate the $k+1$ and k-th iterations of the system. The step-size matrix h is chosen from one of five forms in Table 3.1, depending on the desired stability and numerical performance of the system (see Chapter 4 of this dissertation for a thorough discussion). The vector-valued function f has elements of the form

$$\log\left(\frac{s_j}{\frac{1}{N}\sum_{i=1}^{N} P_{ij}}\right), \quad j = 2, \ldots, J, \tag{3.24}$$

where β is fixed in each P_{ij}. To gain some intuition for (3.23) and (3.24), f can be interpreted as an adjustment term when the step-size matrix is equal to an identity matrix. When $\delta^{(k)}$ results in predicted in-sample market shares that are too small relative to every element in s, then (3.24) is positive and the next value $\delta^{(k+1)}$ is equal to $\delta^{(k)}$ adjusted positively by (3.24). In turn, the positively-adjusted vector $\delta^{(k+1)}$ increases the predicted in-sample market shares relative to the last iteration. On the other hand, negative adjustments are produced when the predicted in-sample market shares are larger than the population shares. Theoretically, these adjustments continue until (3.24) equals zero, or equivalently, until the value of δ that sets the predicted in-sample market shares equal to the population market shares is found. However, because equality in (3.23) may not be possible due to machine precision, the convergence criterion is $\| \delta^{(k+1)} - \delta^{(k)} \| < c$, where c is set to 10^{-14} or smaller following Dube et al. (2011). With the preceding intuition in mind, the different step-size matrices in Table 3.1 are used to accelerate this adjustment process.

The preceding iterative system is nested in gradient-based methods to obtain the MLE for θ. The maximization algorithm searches over the space for β, and for each trial value, (3.23) is used to recover $\delta(\beta)$. The values of β and $\delta(\beta)$ that jointly maximize (3.22) are the maximum likelihood estimates, denoted by $\widehat{\beta}_{MLE}$ and $\widehat{\delta}_{MLE}$. It is important to note that standard maximization algorithms in statistical software (e.g. Matlab, Gauss, etc.) will only output $\widehat{\beta}_{MLE}$ and $\widehat{H}(\widehat{\beta}_{MLE})$, the estimated Hessian matrix evaluated at the maximizing value, because β is the only input to the algorithm. So, to obtain $\widehat{\delta}_{MLE}$, apply the iterative system in (3.23) to $\widehat{\beta}_{MLE}$ after the maximization algorithm has finished. And using a similar approach, the approximate covariance matrix for $\widehat{\theta}_{MLE} = (\widehat{\beta}_{MLE}, \widehat{\delta}_{MLE})$ is obtained numerically by sampling $\beta_{(g)}$ from $\mathcal{N}(\widehat{\beta}_{MLE}, -\widehat{H}(\widehat{\beta}_{MLE})^{-1})$ for $g = 1, 2, \ldots, G$ draws, recovering $\delta(\beta_{(g)})$ from (3.23) for each draw, and computing the sample covariance matrix using the collection of $(\beta_{(g)}, \delta(\beta_{(g)}))$ vectors. This estimated covariance matrix, denoted by $\widehat{T}$, converges to the desired quantity when G is large.

When the market share constraints do not need to be enforced, the MLE for θ is the vector that maximizes (3.9) for observed data with respect to θ. The log-likelihood function is easily maximized with gradient-based algorithms since the analytic score function and Hessian matrix are given in (3.10) and (3.12), respectively. However, caution must be taken during implementation, because the log-likelihood function is not necessarily concave in θ and may have almost-flat spots.

3.5.2 Bayesian estimation

Priors and posterior distribution

For Bayesian analysis, the model from (3.1) to (3.3) is completed with specifications for the prior distributions of the parameters. I will first discuss priors that incorporate the market share information and then describe priors without this information at the end of the section. The prior for θ, denoted by $\pi(\theta|\mu)$, is multivariate normal with mean vector b and covariance matrix B, and it depends on a hyperparameter $\mu = (\mu_1, \mu_2, \ldots, \mu_J)'$ that contains the unknown market shares for all the alternatives in C. Uncertainty for μ is expressed in a hyperprior $\pi(\mu)$. The posterior density of interest with market share information is given by

$$\pi(\theta, \mu|Data) \propto L_B(\theta)\pi(\theta|\mu)\pi(\mu), \tag{3.25}$$

where $Data$ contains the observed broad choice data.

Uncertainty for the market shares is summarized in a hyperprior $\pi(\mu) = \mathcal{D}(\mu|\alpha)$, where

$$\mathcal{D}(\mu|\alpha) = \left(\frac{\Gamma(a_0)}{\prod_{j=1}^{J}\Gamma(\alpha_j)}\right)\prod_{j=1}^{J}(\mu_j)^{\alpha_j - 1} \tag{3.26}$$

is a standardized multivariate Dirichlet density that depends on a vector of parameters $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_J)'$. In (3.26), $a_0 = \sum_{j=1}^{J} \alpha_j$, and each α_j is strictly positive. The Dirichlet prior is convenient for two reasons. First, this prior implies that the market shares are bounded in $(0, 1)$ and must sum to unity. Second, prior information for the market shares can be easily incorporated through α. As an example, suppose that we have prior information in terms of market shares from a previous time period. Then, one approach to incorporate this knowledge into the problem is to set the elements in α such that the prior means are centered around the previous shares and the prior variances reflect our uncertainty about this knowledge. To find these parameters, recall the properties of the Dirichlet distribution: $\mathbb{E}(\mu_j) = \alpha_j/a_0$, and $\mathbb{V}(\mu_j) = \alpha_j(a_0 - \alpha_j)/a_0^2(a_0 + 1)$ for $j = 1, 2, \ldots, J$. Once a value for a_0 is chosen, then the elements of α that set the desired means and variances can be solved for. In general, smaller values for a_0 result in smaller variances.

Information about θ from the market shares is incorporated into $\pi(\theta|\mu)$. I propose specifying the normal prior around the approximate distribution for $\widehat{\theta}_{MLE}$ from a training sample. Specifically, set aside a quarter of the N observations as the training sample, and using only the training sample, estimate θ with the ML methods from the previous section subject to the market share constraints where s is replaced with μ. Then, set the prior mean b equal to $\widehat{\theta}_{MLE}$ and the covariance matrix B equal to $\widehat{T} + \tau I_{(G\times G)}$, where $I_{(G\times G)}$ denotes a $G \times G$ identity matrix. The scalar variance inflation factor τ is used to control the degree of uncertainty in this prior and can also be used for prior sensitivity analysis. The advantages of using this prior specification are discussed next.

Important differences in the Bayesian analysis of the model must be stressed. First, the market shares are treated as unknown quantities, and their uncertainty is reflected in the hyperprior. Being able to express uncertainty in the population shares is important since they are extremely difficult to obtain in practice, and even if they are available, they may be measured with error. For example, in the case of vehicle sales, even “official” market share data are often convoluted with sales to car rental companies, leasing companies, government agencies, and businesses, so the observed shares may not accurately reflect the actual market shares among individual decision makers.

Second, the market share constraints, which relate θ to μ, are not strictly enforced onto the parameters as they are in ML estimation. Instead, the information from μ is incorporated through the prior hyperparameters for θ. Loosely interpreted, with a training sample that is representative of the remaining sample used for inference, the prior for θ is centered around the parameter vector that sets the predicted in-sample market shares equal to μ, and the variance inflation factor τ quantifies our certainty around this vector. As an example, if τ and $\pi(\mu)$ are tightly specified, then our prior knowledge for θ is highly concentrated around the vector that satisfies the constraints implied by μ, which is similar to enforcing the constraints into the model. This uncertainty is important to quantify since the constraints do not need to hold with equality in small sample, and any parameter values that are recovered by strictly enforcing the constraints may be misleading.

For the case without incorporating market share information, let $\pi(\theta)$ be a normal prior for θ with mean vector d and covariance matrix D. The resulting posterior distribution is $\pi(\theta|Data) \propto L_B(\theta)\pi(\theta)$. Note that although the market share information is not incorporated in the prior, we may learn about the parameters as long as the priors for β and δ are dependent.

Markov chain Monte Carlo algorithm

Bayesian estimation is performed by Markov chain Monte Carlo (MCMC) methods with Metropolis-Hastings (MH) steps. Broadly speaking, this method generates samples from the posterior distribution by first proposing candidate values from a known proposal distribution and then accepting them with a specific MH probability. If a proposed candidate value is rejected, then the previous value is used instead. Because the same values may be repeatedly used, this method constructs a Markov chain. After a sufficient burn-in period, the draws from the constructed Markov chain are from the posterior distribution of interest by MH convergence results (Chib and Greenberg, 1995; Tierney, 1994). The final set of posterior draws is then used to construct quantities of interest (e.g. posterior means, standard deviations, etc.).

The MCMC algorithm is described first for the case with market share information. At iteration t, candidate values for β, δ, and μ are proposed as follows

1. Draw $\mu^{(t)}$ from $q_1(\mu) = \mathcal{D}(\mu|\alpha)$.
2. Given $\mu^{(t)}$, maximize $L_B(\theta)\pi(\theta|\mu^{(t)})$ with respect to θ. Denote $\widehat{\theta}$ as the maximizing value and $-\widehat{H}(\widehat{\theta})^{-1}$ as the negative inverse of the estimated Hessian evaluated at $\widehat{\theta}$.
3. Draw $\theta^{(t)}$ from $q_2(\theta|\mu) = f_t(\theta|\widehat{\theta}, -\widehat{H}(\widehat{\theta})^{-1}, \nu)$, which is a t density with location parameter $\widehat{\theta}$, scale matrix $-\widehat{H}(\widehat{\theta})^{-1}$, and degrees of freedom parameter ν, which is set to a small number to ensure heavy tails for this distribution.

The vector $\eta^{(t)} = (\beta^{(t)}, \delta^{(t)}, \mu^{(t)})$ constitutes a proposed candidate draw from the posterior distribution in (3.25), where the proposal density is given by

$$q_3(\eta) \quad = \quad q_1(\mu)q_2(\theta|\mu). \tag{3.27}$$

Upon denoting the right hand side of (3.25) as $p_1(\eta)$, the candidate vector is accepted for iteration t with the MH transition probability

$$\alpha_{\mathrm{MH}}(\eta^{(t-1)}, \eta^{(t)}) = \min\left\{1, \frac{p_1(\eta^{(t)})q_3(\eta^{(t-1)})}{p_1(\eta^{(t-1)})q_3(\eta^{(t)})}\right\}. \tag{3.28}$$

If the candidate vector is not accepted, then treat $\eta^{(t-1)}$ as the draw for iteration t. This iterative process is repeated many times and the final collection of vectors represents the draws from the posterior distribution with market share information.

A few subtle points about the preceding algorithm are now noted. First, the iterative system from (3.23) is nested into each MCMC iteration. It is needed in Step 2 to compute the parameters in $\pi(\theta|\mu^{(t)})$. Despite being nested in the algorithm, the iterative system is inexpensive to evaluate, because the previous draws of η can be used as the starting values to the system. This method significantly reduces the number of iterations needed for the iterative system to converge. Second, if the maximization in Step 2 is difficult perform, then a random walk proposal for $q_2(\theta|\mu)$ is a viable alternative. And third, the candidate vectors for μ are generated from the prior density in (3.26) instead of a density that is proportional to the posterior distribution. This method of generating candidate vectors may be inefficient in the sense that a large portion of the candidates for μ may not be accepted, but its performance is quite good when (3.26) is tightly specified. Otherwise, valid draws of μ from the posterior are incredibly difficult to obtain due to the support restrictions on the market shares.

The MCMC algorithm for the case without market share information is similar. Before starting the algorithm, maximize $p_2(\theta) = L_B(\theta)\pi(\theta)$ with respect to θ, and denote $\widehat{\theta}$ and $-\widehat{H}(\widehat{\theta})^{-1}$ as the maximizing value and negative inverse of the estimated Hessian evaluated at $\widehat{\theta}$, respectively. Then, at step t of the algorithm, candidate vectors for

$\theta^{(t)}$ are proposed from $q_4(\theta) = f_t(\theta|\widehat{\theta}, -\widehat{H}(\widehat{\theta})^{-1}, \nu)$ and accepted with the transition probability

$$\alpha_{\text{MH}}(\theta^{(t-1)}, \theta^{(t)}) = \min\left\{1, \frac{p_2(\theta^{(t)})q_4(\theta^{(t-1)})}{p_2(\theta^{(t-1)})q_4(\theta^{(t)})}\right\}. \tag{3.29}$$

If the candidate vector is not accepted, then set $\theta^{(t)} = \theta^{(t-1)}$.

3.6 Simulation study

This section applies the estimation methods developed in Section 3.5 to simulated data. The results are used to compare the different estimators and to highlight some key points regarding the inclusion of market share information. For this simulation study, the maximum likelihood estimators are analyzed in a repeated sampling study, and the Bayesian methods are analyzed using a single data set from the sampling study.

The population data for 20000 decision makers are generated based on (3.1) to (3.3). The data-generating values for θ are presented in Table 3.2, and the exogenous attributes for all decision makers and alternatives are independently drawn from a standard normal distribution. For these decision makers, they are faced with ten alternatives in the choice set, and the broad groups are defined as $C_1 = \{1, 2, \ldots, 9\}$ and $C_2 = \{10\}$. Note that these group configurations are constructed so that the observed broad choice data are fairly uninformative for θ. With one exogenous attribute for each decision maker and alternative, there are ten parameters to estimate: $\theta = (\delta_2, \delta_3, \ldots, \delta_{10}, \beta)'$.

		ECD		BCD		BCDC	
θ	True	Est.	SE	Est.	SE	Est.	SE
δ_2	-0.68	-0.68	0.08	-0.72	0.74	-0.69	0.02
δ_3	0.87	0.87	0.07	0.86	0.57	0.91	0.03
δ_4	1.89	1.89	0.06	1.93	0.52	1.91	0.06
δ_5	4.13	4.14	0.07	4.18	0.49	4.09	0.13
δ_6	1.16	1.17	0.07	1.19	0.56	1.10	0.03
δ_7	1.65	1.65	0.06	1.68	0.54	1.69	0.05
δ_8	1.39	1.40	0.06	1.42	0.54	1.34	0.04
δ_9	2.21	2.21	0.06	2.24	0.53	2.25	0.07
δ_{10}	-1.01	-1.01	0.08	-1.00	0.43	-1.00	0.03
β	2.98	2.98	0.03	3.01	0.11	2.99	0.11

Table 3.2: Maximum likelihood estimates of θ using exact choice data (ECD), broad choice data (BCD), and broad choice data with constraints (BCDC).

The repeated sampling study estimates θ over 1000 repetitions. For each repetition, a subset of 15000 decision makers is randomly sampled from the population, and θ is estimated with the ML methods from Section 3.5.1 based on exact choice data (ECD), broad choice data (BCD), and broad choice data with the market share constraints enforced (BCDC). The population market shares that are used in BCDC and throughout this section are calculated from the full population and remain constant over each repetition. The population shares for alternatives one through ten are roughly 5%, 3%, 7%, 12%, 28%, 8%, 11%, 9%, 14%, and 3%, respectively. In general, the in-sample market shares from each repetition closely mimic the ones from the population.

The ML estimates corresponding to θ are presented in Table 3.2. Comparing the numerical results between ECD and BCD, both sets of estimates are close to their true values, but there is substantially more variability when BCD is used instead of ECD. In particular, the standard errors corresponding to β and δ roughly differ by factors of four and eight, respectively. The difference in variability is expected since there is generally less information in the broad choice data than in the exact choice data. In addition, this difference is amplified by the specified group configurations.

Figure 3.3 illustrates the sampling distributions for the estimators of δ_2 and β (the distributions for the other estimators are omitted since they are qualitatively similar to the one for δ_2) and confirms that the distributions based on BCD are wider than the ones based on ECD.

With the population market share constraints enforced in BCDC, the resulting ML estimates are generally close to their true parameter values. In terms of variability, the standard errors that correspond to δ are substantially smaller than the errors obtained with ECD and BCD. This suggests that the constraints are helpful in pinning down the estimates of δ. On the other hand, the standard error corresponding to β does not differ from one obtained using BCD, which is around four times larger than the one from ECD. This suggests that the constraints do not contain much information with regards to β. Figure 3.3 confirms the first observation as the sampling distribution corresponding to δ_2 is highly concentrated around the true value of -0.68 when using BCDC, even more so than the ones for ECD and BCD. The same figure also confirms that the distributions corresponding to β are almost identical between ECD and BCDC.

Bayesian estimation is based on a data set from a single repetition of the repeated sampling study. Similar to the ML discussion, I present the posterior means and standard deviations for θ across three cases. The first two cases respectively correspond to ECD and BCD with priors that are fairly non-informative. For these priors, I set $b = 0_{(G\times 1)}$ and $B = 1000 \times I_{(G\times G)}$. The remaining case is based on broad choice data with an informative prior for θ (BCDIP). As discussed in Section 3.5.2, this prior is developed using the remaining 5000 observations as a training sample and the population market shares. The hyperparameter τ is set to 0.1 which slightly inflates the prior variances for θ. Note that the known population market shares are used in this simulation exercise, so μ is known with certainty. And as a result, the

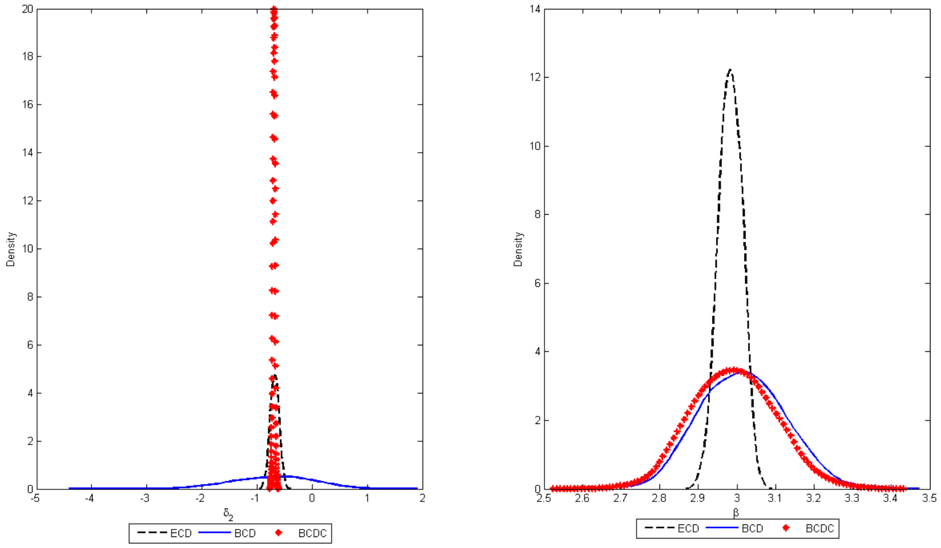

Figure 3.3: Sampling distributions for the maximum likelihood estimators of δ_2 and β. These are based on exact choice data (ECD), broad choice data (BCD), and broad choice data with population market share constraints (BCDC).

hyperprior $\pi(\mu)$ does not need to be specified. The prior values for b and the standard deviations implied by B are in Table 3.3. The off-diagonals of B are close to zero and are not reported.

Table 3.3 contains the Bayesian estimates based on 10000 runs of the MCMC algorithm with 2000 runs discarded for burn-in. The numerical results corresponding to ECD are similar to those from ML estimation. That is, when ECD is used, the posterior means are quite close to the true values of the parameters used to generate the data, and the standard deviations are fairly tight. This suggests that the exact choice data are very informative about the parameters. For the case with BCD, the posterior means are in the neighborhood of the true values, but the posterior standard deviations are relatively large compared to the ones from ECD. Similar to the ML study, the posterior standard deviations for β and δ differ by factors of four and eight, respectively. The posterior distributions depicted in Figure 3.4 confirm these

		ECD		BCD		BCDIP		BCDIP Prior	
θ	True	Mean	SD	Mean	SD	Mean	SD	Mean	SD
δ_2	-0.68	-0.69	0.08	-0.68	0.72	-0.76	0.27	-0.75	0.32
δ_3	0.87	0.92	0.07	0.73	0.60	0.75	0.23	0.89	0.32
δ_4	1.89	1.98	0.06	2.33	0.49	1.98	0.23	1.85	0.33
δ_5	4.13	4.18	0.07	4.29	0.48	4.07	0.22	4.19	0.39
δ_6	1.16	1.17	0.07	1.18	0.51	1.07	0.24	1.18	0.32
δ_7	1.65	1.75	0.07	1.64	0.56	1.46	0.24	1.56	0.33
δ_8	1.39	1.41	0.07	1.72	0.53	1.38	0.25	1.29	0.32
δ_9	2.21	2.33	0.07	2.65	0.51	2.34	0.22	2.24	0.34
δ_{10}	-1.01	-0.94	0.08	-0.82	0.43	-1.05	0.14	-1.20	0.32
β	2.98	3.02	0.03	3.03	0.11	3.03	0.10	3.09	0.37

Table 3.3: Posterior estimates for θ. These are based on exact choice data (ECD), broad choice data (BCD), and broad choice data with an informative prior (BCDIP).

observations.

In the case of BCDIP, the posterior means for δ are generally closer to the true values than the ones obtained using BCD, and the standard deviations are relatively smaller. For β, the posterior estimates do not change much over the ones obtained using BCD, which is a similar conclusion to the ML study. These observations are confirmed by Figure 3.4. From the figure, we see that the posterior distribution for δ_2 has less variability in this case relative to BCD. Also, note that this distribution for δ_2 is not as highly concentrated as the one from ML estimation. This additional variability comes from the prior which expresses some uncertainty in the constraints. For β, the figure confirms that the posterior distributions are almost identical. These Bayesian estimates suggest two conclusions. The first one is that the posterior distribution for δ is sensitive to the prior distribution despite the large sample size. This is expected since the broad choice data in this simulation exercise are constructed to be uninformative about θ. The second conclusion is that the broad choice data are informative for β, because the posterior estimates between BCD and BCDIP do not change much despite a fairly informative prior used in BCDIP.

A few conclusions for the entire simulation study are in order. First, the market shares in the form of constraints or prior information contain more information about δ than β. Second, the addition of market share information improves both estimation methods over the case without incorporating this information. However, I must warn that the effectiveness of this technique critically depends on the quality of the market shares and the sample size. If the population shares do not closely mimic the predicted in-sample shares, then inference is questionable. This issue is especially problematic for ML estimation, because the parameters are recovered using strict constraints based on the shares. In contrast, for the Bayesian methods, a large value for τ can be used to express uncertainty in this technique. At worst, the resulting prior is relatively non-informative, and we obtain results similar to the Bayesian estimates with BCD. Lastly, for some extreme group configurations, the broad choice data will not be informative for θ (especially δ). For these cases, the Bayesian estimates are highly sensitive to the prior. But despite this strong dependence, the numerical results demonstrate that, as long as the practitioner is thoughtful in forming the prior, the results are well behaved.

3.7 Concluding remarks

This paper introduces a new discrete choice model to analyze choice outcomes that only broadly represent the actual choices made by the decision makers. It is useful in analyzing situations where the choice behavior at a lower level is desired but only higher level choice data are available. The parameters from the proposed model are locally identified, but in some perverse yet interesting cases, they may only be weakly identified. To efficiently recover the parameter estimates in these troublesome cases, I show how population-level market shares can be introduced as additional

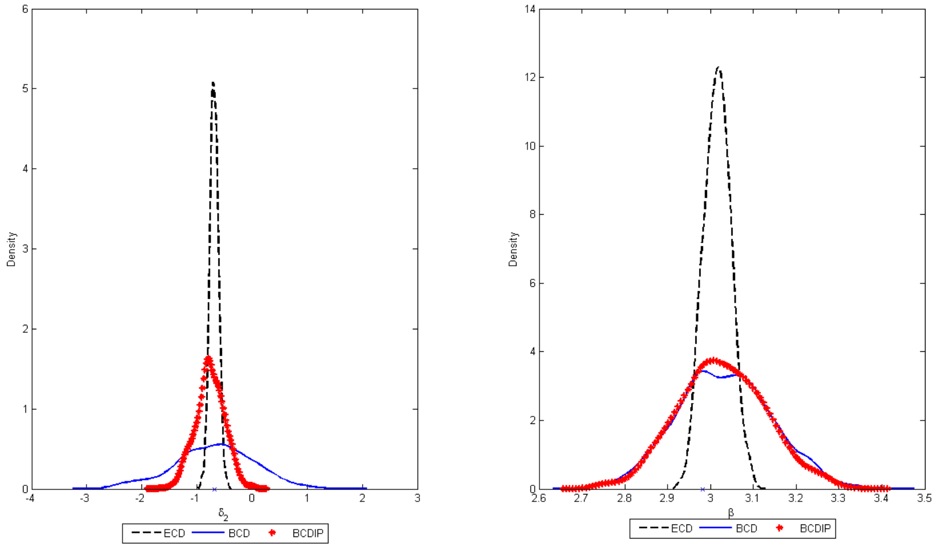

Figure 3.4: Posterior distributions for δ_2 and β. These are based on exact choice data (ECD), broad choice data (BCD), and broad choice data with an informative prior (BCDIP).

information into the problem. A simulation study shows that both maximum likelihood and Bayesian estimation techniques benefit from the inclusion of the market share information. Although the effectiveness of this approach depends critically on the quality of the population market shares, the results demonstrate that meaningful relationships can be uncovered using this new class of models.

3.8 Appendix

3.8.1 Likelihood quantities and the information matrix

This section derives the score function and Hessian matrix of the log-likelihood function for the model with broad choice data. To obtain the score function, expand (3.5) in terms of (3.8), resulting in the log-likelihood function

$$L_B(\theta) = \sum_{i=1}^{N}\sum_{m=1}^{M} Y_{im}\left(\log(\sum_{c\in C_m}\exp(w'_{ic}\theta)) - \log(\sum_{j=1}^{J}\exp(w'_{ij}\theta))\right). \tag{3.30}$$

The score function is then

$$\begin{aligned}
S_B(\theta) &= \frac{\partial L_B(\theta)}{\partial\theta} \tag{3.31}\\
&= \sum_{i=1}^{N}\sum_{m=1}^{M} Y_{im}\left(\frac{\partial}{\partial\theta}\log(\sum_{c\in C_m}\exp(w'_{ic}\theta)) - \frac{\partial}{\partial\theta}\log(\sum_{j=1}^{J}\exp(w'_{ij}\theta))\right) \tag{3.32}\\
&= \sum_{i=1}^{N}\sum_{m=1}^{M} Y_{im}\left(\sum_{c\in C_m} w_{ic}\frac{\exp(w'_{ic}\theta)}{\sum_{s\in C_m}\exp(w'_{is}\theta)} - \sum_{j=1}^{J} w_{ij}\frac{\exp(w'_{ij}\theta)}{\sum_{r=1}^{J}\exp(w'_{ir}\theta)}\right) \tag{3.33}\\
&= \sum_{i=1}^{N}\sum_{m=1}^{M} Y_{im}\left(\sum_{c\in C_m} w_{ic}P_{ic|C_m} - \sum_{j=1}^{J} w_{ij}P_{ij}\right) \tag{3.34}\\
&= \sum_{i=1}^{N}\left(\sum_{m=1}^{M} Y_{im}\sum_{c\in C_m} w_{ic}P_{ic|C_m} - \sum_{m=1}^{M} Y_{im}\sum_{j=1}^{J} w_{ij}P_{ij}\right) \tag{3.35}\\
&= \sum_{i=1}^{N}\left(\sum_{m=1}^{M} Y_{im}\sum_{c\in C_m} w_{ic}P_{ic|C_m} - \sum_{j=1}^{J} w_{ij}P_{ij}(\sum_{m=1}^{M} Y_{im})\right) \tag{3.36}\\
&= \sum_{i=1}^{N}\left(\sum_{m=1}^{M} Y_{im}\sum_{c\in C_m} w_{ic}P_{ic|C_m} - \sum_{j=1}^{J} w_{ij}P_{ij}\right), \tag{3.37}
\end{aligned}$$

since $\sum_{m=1}^{M} Y_{im} = 1$ for all decision makers.

Before deriving the Hessian matrix, a few miscellaneous quantities are needed. Note that

$$\frac{\partial P_{ij}}{\partial\theta'} = P_{ij}(w'_{ij} - \sum_{r=1}^{J} w'_{ir}P_{ir}), \tag{3.38}$$

and

$$\frac{\partial P_{ic|C_m}}{\partial \theta'} = P_{ic|C_m}(w'_{ic} - \sum_{c \in C_m} w'_{ic} P_{ic|C_m}), \tag{3.39}$$

Also,

$$\begin{aligned}
&\sum_{j=1}^{J} w_{ij} P_{ij} w'_{ij} - (\sum_{j=1}^{J} w_{ij} P_{ij})(\sum_{j=1}^{J} w_{ij} P_{ij})' \\
&\qquad = \sum_{j=1}^{J} (w_{ij} - \sum_{r=1}^{J} w_{ir} P_{ir}) P_{ij} (w_{ij} - \sum_{r=1}^{J} w_{ir} P_{ir})',
\end{aligned} \tag{3.40}$$

which can be shown easily by adding and subtracting $(\sum_{r=1}^{J} w_{ir} P_{ir})(\sum_{r=1}^{J} w_{ir} P_{ir})'$ to the left hand side of (3.40) and manipulating the summation indices. And similarly,

$$\begin{aligned}
&\sum_{c \in C_m} w_{ic} P_{ic|C_m} w'_{ic} - (\sum_{c \in C_m} w_{ic} P_{ic|C_m})(\sum_{c \in C_m} w_{ic} P_{ic|C_m})' \\
&\qquad = \sum_{c \in C_m} (w_{ic} - \sum_{s \in C_m} w_{is} P_{is|C_m}) P_{ic|C_m} (w_{ic} - \sum_{s \in C_m} w_{is} P_{is|C_m})'.
\end{aligned} \tag{3.41}$$

The Hessian matrix is

$$\begin{aligned}
H_B(\theta) &= \frac{\partial S_B(\theta)}{\partial \theta'} && (3.42) \\
&= \sum_{i=1}^{N} \left(\sum_{m=1}^{M} Y_{im} \sum_{c \in C_m} w_{ic} \frac{\partial P_{ic|C_m}}{\partial \theta'} - \sum_{j=1}^{J} w_{ij} \frac{\partial P_{ij}}{\partial \theta'} \right) && (3.43) \\
&= \left(\sum_{i=1}^{N} \sum_{m=1}^{M} Y_{im} \sum_{c \in C_m} w_{ic} \frac{\partial P_{ic|C_m}}{\partial \theta'} \right) - \left(\sum_{i=1}^{N} \sum_{j=1}^{J} w_{ij} \frac{\partial P_{ij}}{\partial \theta'} \right) && (3.44) \\
&= L - F, && (3.45)
\end{aligned}$$

because

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{m=1}^{M} Y_{im} \sum_{c\in C_m} w_{ic} \frac{\partial P_{ic|C_m}}{\partial \theta'} && (3.46)\\
&= \sum_{i=1}^{N}\sum_{m=1}^{M} Y_{im} \sum_{c\in C_m} w_{ic} P_{ic|C_m} (w'_{ic} - \sum_{c\in C_m} w'_{ic} P_{ic|C_m}) && (3.47)\\
&= \sum_{i=1}^{N}\sum_{m=1}^{M} Y_{im} \sum_{c\in C_m} (w_{ic} P_{ic|C_m} w'_{ic} - w_{ic} P_{ic|C_m} \sum_{c\in C_m} w'_{ic} P_{ic|C_m}) && (3.48)\\
&= \sum_{i=1}^{N}\sum_{m=1}^{M} Y_{im} (\sum_{c\in C_m} w_{ic} P_{ic|C_m} w'_{ic} - (\sum_{c\in C_m} w_{ic} P_{ic|C_m})(\sum_{c\in C_m} w_{ic} P_{ic|C_m})') && (3.49)\\
&= \sum_{i=1}^{N}\sum_{m=1}^{M} Y_{im} \sum_{c\in C_m} (w_{ic} - \sum_{s\in C_m} w_{is} P_{is|C_m}) P_{ic|C_m} (w_{ic} - \sum_{s\in C_m} w_{is} P_{is|C_m})' && (3.50)\\
&= L. && (3.51)
\end{aligned}$$

where moving from (3.46) to (3.47) uses the expression from (3.39), and moving from (3.49) to (3.50) uses (3.41). Also, based on a similar argument, $\sum_{i=1}^{N}\sum_{j=1}^{J} w_{ij} \frac{\partial P_{ij}}{\partial \theta'} = F$.

The information matrix is easy to derive using the preceding quantities. Note that the only random terms in $H_B(\theta)$ are Y_{im} and that $\mathbb{E}(Y_{im}) = \tilde{P}_{im}$ for all decision makers and groups. Plugging the preceding expectations into $-\mathbb{E}(H_B(\theta))$ results in the desired quantity in (3.15).

3.8.2 Forms for $h(\delta^{(k)})$

The different forms for $h(\delta^{(k)})$ in the iterative system $\delta^{(k+1)} = \delta^{(k)} + h(\delta^{(k)}) f(\delta^{(k)})$ are

1. $h_{BLP}(\delta^{(k)}) = I_{(J-1\times J-1)}$. This step size results in the contraction mapping

algorithm from Berry et al. (1995).

2. $h_{ANJ}(\delta^{(k)}) = -[f'(\delta^{(k)})]^{-1}$, where f' is the Jacobian matrix. This is the standard analytic Newton Jacobian (ANJ) step size and is generally not a diagonal matrix.

3. $h_{APNJ}(\delta^{(k)}) = -[a(s)]^{-1}$, where a is an approximation to the Jacobian matrix that depends on the known market shares. A specific form for a is given at the end. This is the approximated Newton Jacobian (APNJ) step size and is also not a diagonal matrix.

4. $h_{ADJ}(\delta^{(k)}) = -Diag(f'(\delta^{(k)}))^{-1}$. This step size is referred to as the analytic diagonal Jacobian (ADJ). This specification results in a diagonal matrix with entries equal to the negative reciprocals of the diagonal elements from the Jacobian. As an illustration, when there are two fixed points, $\delta = (\delta_2, \delta_3)$, then

$$f'(\delta) = \begin{pmatrix} \frac{\partial f_1}{\partial \delta_2} & \frac{\partial f_1}{\partial \delta_3} \\ \frac{\partial f_2}{\partial \delta_2} & \frac{\partial f_2}{\partial \delta_3} \end{pmatrix}, \quad \text{and} \quad h_{ADJ}(\delta) = \begin{pmatrix} -(\frac{\partial f_1}{\partial \delta_2})^{-1} & 0 \\ 0 & -(\frac{\partial f_2}{\partial \delta_3})^{-1} \end{pmatrix}.$$

5. $h_{APDJ}(\delta^{(k)}) = -Diag(a(s))^{-1}$. This specification is based on ADJ but uses the approximation to the Jacobian matrix (referred to as approximated diagonal Jacobian (APDJ)). Similar to ADJ, the matrix resulting from APDJ is diagonal with entries that equal the negative reciprocals of the diagonal elements of $a(s)$.

The Jacobian matrix for f can be shown to equal

$$f'(\delta) = \begin{pmatrix} 1 - \frac{\sum_{i=1}^N P_{i2}^2}{\sum_{i=1}^N P_{i2}} & -\frac{\sum_{i=1}^N P_{i2}P_{i3}}{\sum_{i=1}^N P_{i2}} & \dots & -\frac{\sum_{i=1}^N P_{i2}P_{iJ}}{\sum_{i=1}^N P_{i2}} \\ -\frac{\sum_{i=1}^N P_{i3}P_{i2}}{\sum_{i=1}^N P_{i3}} & 1 - \frac{\sum_{i=1}^N P_{i3}^2}{\sum_{i=1}^N P_{i3}} & \dots & -\frac{\sum_{i=1}^N P_{i3}P_{iJ}}{\sum_{i=1}^N P_{i3}} \\ \vdots & \dots & \dots & \vdots \\ -\frac{\sum_{i=1}^N P_{iJ}P_{i2}}{\sum_{i=1}^N P_{iJ}} & -\frac{\sum_{i=1}^N P_{iJ}P_{i3}}{\sum_{i=1}^N P_{iJ}} & \dots & 1 - \frac{\sum_{i=1}^N P_{iJ}^2}{\sum_{i=1}^N P_{iJ}} \end{pmatrix}. \tag{3.52}$$

The step sizes $h_{ADJ}(\delta^{(k)})$ and $h_{ANJ}(\delta^{(k)})$ are constructed using (4.6).

The approximated step sizes $h_{APDJ}(\delta^{(k)})$ and $h_{APNJ}(\delta^{(k)})$ are based on known market shares around the fixed point and the assumption of homogenous decision makers. Specifically, from the market share constraints, we know that $s_j = N^{-1}\sum_{i=1}^{N} P_{ij}$ must hold for $j = 2, 3, \ldots, J$ around the fixed point, therefore all the denominators in (4.6) of the form $\sum_{i=1}^{N} P_{ij}$ are approximated with $s_j \times N$. Next, by assuming that each decision maker behaves like the market shares, we can approximate $\sum_{i=1}^{N} P_{ij}P_{ik}$ from (4.6) with $s_j \times s_k \times N$. With these assumptions, $f'(\delta)$ is approximated as

$$a(s) = \begin{pmatrix} 1 - s_2 & -s_3 & \ldots & -s_J \\ -s_2 & 1 - s_3 & \ldots & -s_J \\ \vdots & \ldots & \ldots & \vdots \\ -s_2 & -s_3 & \ldots & 1 - s_J \end{pmatrix}. \tag{3.53}$$

This approximated matrix is used to construct $h_{APDJ}(\delta^{(k)})$ and $h_{APNJ}(\delta^{(k)})$. The approximations are better when the iterative system is close to the fixed point and when the decision makers behave like the market shares. Even if these assumptions do not hold exactly, the resulting iterative systems should still be faster than the BLP system. Another feature of this approximation is that it does not depend on δ. Unlike the analytic versions, it is not necessary to recalculate these step sizes at each δ or in each iteration of the system. This saves a lot of time because multiple evaluations of the partial derivatives and inversions matrices are avoided. The biggest improvements occur when J is large.

Chapter 4

Speeding up the BLP contraction mapping

4.1 Introduction

The model from Berry et al. (1995) (hereafter referred to as BLP) is widely used in applied work. It is a key tool to examine questions regarding market power, mergers, innovation, and valuation of new brands in differentiated-product industries. The main estimation algorithm involves minimizing a generalized method of moments (GMM) objective function which depends on the solution to a high-dimensional system of nonlinear equations. To solve this system of equations, Berry (1994) and BLP suggest using a contraction mapping. The mapping is implemented as a multivariate iterative system that, when iterated enough times, will converge within a certain criteria to a unique vector of fixed points. This iterative system is solved every time the GMM objective function is evaluated in a minimization routine.

The BLP model has been estimated in a variety of ways. For general overviews, see

Knittel and Metaxoglou (2012) and Nevo (2000). From the theoretical side, Jiang et al. (2009), Musalem et al. (2009), and Yang et al. (2003) utilize Bayesian Markov chain Monte Carlo (MCMC) estimation algorithms, and Park and Gupta (2009, 2011) provide simulated maximum likelihood (MSL) estimators. Most of the aforementioned papers also embed the BLP contraction mapping inside their proposed estimation procedures which is by far the most popular approach.

A growing body of research is concerned with how the convergence of the inner BLP-style contraction mapping affects the outer estimation routines. Dube et al. (2011), Judd and Skrainka (2011), Knittel and Metaxoglou (2012), and Skrainka (2012) warn that estimating the unknown parameters in the BLP model with a combination of an outer non-linear search algorithm (e.g. the GMM minimization or MSL maximization algorithm) and the inner contraction mapping can be difficult. In particular, Dube et al. (2011) conclude that errors from the inner contraction mapping will propagate to the outer objective function and result in failures to converge. These errors are results from incorrectly specifying the convergence criteria to be too loose for the inner contraction mapping. This practice of setting loose convergence criteria is not uncommon in applied work to reduce computational demands of the estimation procedures. Although this warning is in the context of GMM estimation, intuitively it should hold for any other method of estimation that requires the contraction mapping. Therefore, as a rule of thumb, Dube et al. (2011) recommend setting the inner convergence criteria to 10^{-14} or smaller.

Setting such a tight convergence criteria for the inner iterative system is a major burden in empirical work. While it is relatively quick for the iterative system to converge for a single run of the algorithm, it is extremely expensive in terms of computational time when repeated evaluations are needed, particularly when the algorithm is embedded into the aforementioned estimation routines. It is common to evaluate the contraction mapping at least a few hundred times during estimation. To lighten this computational burden, Dube et al. (2011) recommend imposing the system of nonlinear equations as constraints and using constrained maximum likelihood to estimate the unknown parameters. They find that estimation is an order of magnitude faster than using the contraction mapping, but this approach is still slow when there are many alternatives across the markets. Besides this paper, I am unaware of any other work in the literature on explicitly speeding up the BLP contraction mapping.

In this paper, I explore four simple modifications for the BLP contraction mapping to improve its rate of convergence. The modifications differ in their tradeoffs between convergence speed, numerical stability, and ease of implementation, so it is up to the practitioner to choose the most suitable modification for his problem. The motivation for these modifications come from the relationship between contraction mappings and fixed-point algorithms. In particular, I employ both analytic and approximated Newton algorithms (also known as Newton-Raphson algorithms) to solve the system of nonlinear equations. A unique contribution of this paper is in the approximated Newton step size which exploits knowledge of the true market shares close to the fixed points. An important implication of this approximation is that this step size does not need to be recalculated at each parameter value or in each iteration of the iterative system. This feature significantly reduces computational time, and as the simulation study will show, improves the numerical stability of the iterative system relative to the analytic Newton algorithm.

Moreover, the modifications are designed such that they can be incorporated into existing BLP-style contraction mapping code with minimal effort. I suspect that a majority of the applied researchers using the BLP model have the original contraction mapping specification coded already, so in order to reduce the amount of additional coding, I have kept the modifications relatively simple. While the modifications proposed in this paper may not converge as fast as the "cutting-edge" root-finding algorithms, my modifications are simple to code (typically only a few additional lines of code are needed), and the practitioner does not need to worry about whether their statistical software (e.g. R, Matlab, Gauss, etc.) has these "cutting-edge" algorithms.

In a simulation study, I show that the new algorithms require significantly fewer iterations to converge to the unique vector of fixed points relative to the original BLP specification. The "best" modification is the one based on the analytic Newton step and results in an 80-fold improvement. The approximated Newton steps also perform well and result in an 8-fold improvement.

The remainder of the paper is organized as follows. Section 4.2 provides the necessary background on the theory of fixed-point algorithms and contraction mappings. Based on the intuition from the fixed-point algorithms, Section 4.3 describes the four proposed modifications, while Section 4.4 demonstrates their performance in a simulation study. The conclusion and directions for future study are contained in Section 4.5.

4.2 Background

The following definitions and theorems are all from Olver (2006).

Definition 2 *An iterative system has the form* $\delta^{(k+1)} = g(\delta^{(k)})$*, where* $g : \mathbb{R}^J \to \mathbb{R}^J$ *is a real vector-valued function.*

Note that the (k) and $(k+1)$ superscripts in the preceding definitions respectively denote the k-th and $(k+1)$-th iterations of the iterative system and not higher order derivatives.

Definition 3 *A fixed point to an iterative system is a vector $\delta^* \in \mathbb{R}^J$ such that $g(\delta^*) = \delta^*$.*

Definition 4 *A function $g : \mathbb{R}^J \rightarrow \mathbb{R}^J$ is called a contraction mapping on a domain $\Omega \subset \mathbb{R}^J$ if*

1. *it maps Ω to itself, so $g(\delta) \in \Omega$ whenever $\delta \in \Omega$, and*
2. *there exists a constant $0 \leq \sigma < 1$ such that*

$$\parallel g(\delta_0) - g(\delta_1) \parallel \leq \sigma \parallel \delta_0 - \delta_1 \parallel \quad \textit{for all} \quad \delta_0, \delta_1 \in \Omega. \tag{4.1}$$

Lemma 1 *If $g : \Omega \rightarrow \Omega$ and $\parallel g'(\delta) \parallel < 1$ for all $\delta \in \Omega$, then g is a contraction mapping.*

Theorem 1 *If g is a contraction mapping on a bounded domain $\Omega \subset \mathbb{R}^J$, then g admits a unique fixed point δ^*. Moreover, starting with any initial point $\delta^{(0)} \in \Omega$, the iterates $\delta^{(k+1)} = g(\delta^{(k)})$ necessarily converge to the fixed point: $\delta^{(k)} \rightarrow \delta^*$.*

The BLP iterative system has the form $\delta^{(k+1)} = g(\delta^{(k)})$, where the j-th value of the vector-valued function g is $g_j(\delta^{(k)}) = \delta_j^{(k)} + \log(s_j / N^{-1} \sum_{i=1}^{N} P_{ij}(\delta^{(k)}))$ (more details on this iterative system later). BLP proves that this is a contraction mapping (Berry et al., 1995), but it can also be shown easily by verifying Lemma 1 for any matrix norm (e.g. the p, 1, or ∞-norms). The advantage to this iterative system is that we are guaranteed by Theorem 1 to converge to the unique fixed point regardless of the starting value. However, the disadvantage is that this process only has a linear rate of convergence.

In general, the matrix-norm of $g'(\delta)$ from Lemma 1 governs the rate of convergence for the contraction mapping, where the mapping converges fastest when $g'(\delta) = 0_{(J\times J)}$. Intuitively, this says that we move across the iterative function the fastest when it is flat in all directions. This intuition is used to construct an iterative system that can converge very rapidly.

Suppose that we want to design an efficient fixed-point iterative system $\delta^{(k+1)} = g(\delta^{(k)}) = \delta^{(k)} + h(\delta^{(k)})f(\delta^{(k)})$ that has a fixed point at δ^* whenever $f(\delta^*) = 0$. Then, following the intuition from the preceding paragraph, the fastest convergence occurs whenever $g'(\delta) = 0_{(J\times J)}$, so the problem boils down to finding $h(\delta^{(k)})$ such that this condition is satisfied. The solution is $h(\delta^{(k)}) = -[f'(\delta^{(k)})]^{-1}$ and results in the Newton iterative system.

Definition 5 *The Newton iterative system is* $\delta^{(k+1)} = \delta^{(k)} - [f'(\delta^{(k)})]^{-1}f(\delta^{(k)})$*, where* $f'(\cdot)$ *is the Jacobian of* f*, and the vector-valued function* f *equals a vector of zeros* $0_{J\times 1}$ *at the fixed point (i.e.* $f(\delta^*) = 0_{(J\times 1)}$*).*

Theorem 2 *Let* δ^* *be a solution to the system* $f(\delta^*) = 0$*. Then, provided* $\delta^{(0)}$ *is sufficiently close to* δ^**, the Newton iteration scheme converges at a quadratic rate to the solution* δ^**.*

Using similar notation, the BLP iterative system can be written as $\delta^{(k+1)} = \delta^{(k)} + I_{(J\times J)}f(\delta^{(k)})$, where $I_{(J\times J)}$ is a $J \times J$ identity matrix, and the vector-valued function f contains elements of the form $f_j(\delta_j^{(k)}) = \log(s_j/N^{-1}\sum_{i=1}^{N} P_{ij}(\delta^{(k)}))$. Clearly, the vector δ that sets the predicted market shares equal to the known market shares is a fixed point of this iterative system, but because $I_{(J\times J)} \neq -[f'(\delta^{(k)})]^{-1}$, the original BLP contraction mapping does not converge as fast as the Newton iterative system near the fixed point.

Therefore, the BLP rate of convergence can be improved with Newton-type algorithms or variations of it. It is important to note that the BLP contraction mapping is globally convergent by Theorem 1, while the Newton algorithms are convergent around the fixed point. Provided that we have reasonable starting values, the locally convergent nature of the Newton algorithms should not be an issue. In the next section, I will explore four different specifications for $h(\delta^{(k)})$.

4.3 Four proposals for $h(\delta^{(k)})$

Using the framework from the preceding section, the BLP iterative system and the proposed modifications in this paper can be expressed and differentiated by their expressions for $h(\delta^{(k)})$ (referred to as step sizes) in the iterative system $\delta^{(k+1)} = g(\delta^{(k)}) = \delta^{(k)} + h(\delta^{(k)})f(\delta^{(k)})$. They are

1. $h_{BLP}(\delta^{(k)}) = I_{(J\times J)}$. This step size results in the standard BLP contraction mapping.

2. $h_{ANJ}(\delta^{(k)}) = -[f'(\delta^{(k)})]^{-1}$. This is the standard analytic Newton Jacobian (ANJ) and is generally not a diagonal matrix.

3. $h_{APNJ}(\delta^{(k)}) = -[a(s)]^{-1}$. This is the approximated Newton Jacobian (APNJ) and is also not a diagonal matrix. The approximation function a is based on the known market shares, denoted as a vector s, around the fixed point. A specific example will be given in the subsequent discussion.

4. $h_{ADJ}(\delta^{(k)}) = -Diag(f'(\delta^{(k)}))^{-1}$. This step size is referred to as the analytic diagonal Jacobian (ADJ). This specification results in a diagonal matrix with entries as the negative reciprocals of the diagonal elements from the Jacobian $f'(\delta^{(k)})$. As an illustration, when there are two fixed points, $\delta = (\delta_1, \delta_2)$, then

$$f'(\delta) = \begin{pmatrix} \frac{\partial f_1}{\partial \delta_1} & \frac{\partial f_1}{\partial \delta_2} \\ \frac{\partial f_2}{\partial \delta_1} & \frac{\partial f_2}{\partial \delta_2} \end{pmatrix}, \quad \text{and} \quad h_{ADJ}(\delta) = \begin{pmatrix} -(\frac{\partial f_1}{\partial \delta_1})^{-1} & 0 \\ 0 & -(\frac{\partial f_2}{\partial \delta_2})^{-1} \end{pmatrix}.$$

5. $h_{APDJ}(\delta^{(k)}) = -Diag(a(s))^{-1}$. This specification is based on ADJ but uses the approximation to the Jacobian matrix $f'(\delta^{(k)})$ (referred to as approximated diagonal Jacobian (APDJ)). Similar to ADJ, the matrix resulting from APDJ is diagonal with entries that equal the negative reciprocals of the diagonal elements of $a(s)$.

Summaries of the methods are in Tables 4.1 and 4.2. The motivation for the form of $h_{ANJ}(\delta^{(k)})$ was discussed in the previous section: it sets the iterative function flat in all directions with respect to all variables which results in rapid convergence rates. This form for the step size should result in the fastest iterative system. However, this matrix can be difficult to compute when Jacobian is not in a nice form, and more importantly, difficult to invert numerically when the Jacobian is numerically unstable. To avoid the computational obstacle of $h_{ANJ}(\delta^{(k)})$, the approximated version, $h_{APNJ}(\delta^{(k)})$, is a viable alternative. This alternative step size shares the same intuition and should have the second fastest rate of convergence.

To overcome the invertibility issue of $h_{ANJ}(\delta^{(k)})$, $h_{ADJ}(\delta^{(k)})$ can be used instead. It still sets the iterative function g flat in certain directions but not with respect to all the variables (the off-diagonals of g' will not necessarily equal 0). This form is easily invertible due to the diagonal structure and should offer moderate improvements in convergence speed. However, in the case that $h_{ADJ}(\delta^{(k)})$ is difficult to compute, we can resort to the approximation of it: $h_{APDJ}(\delta^{(k)})$. This form is also easy to invert and has the added advantage of being easy to compute due to the approximation (more details later).

Method	Form of $h(\delta^{(k)})$
h_{BLP}	$I_{J\times J}$
h_{ANJ}	$-[f'(\delta^{(k)})]^{-1}$
h_{APNJ}	$-[a(s)]^{-1}$
h_{ADJ}	$-Diag(f'(\delta^{(k)}))^{-1}$
h_{APDJ}	$-Diag(a(s))^{-1}$

Table 4.1: Forms for the various step size matrices.

Method	Pros	Cons
h_{BLP}	Global convergence	Linear convergence
h_{ANJ}	Quadratic convergence	Hard to compute, invert
h_{APNJ}	Easy to compute, stable	Hard to invert
h_{ADJ}	Easy to invert, stable	Hard to compute, inefficient
h_{APDJ}	Easy to compute, invert, stable	Inefficient

Table 4.2: Pros and cons of the various methods. In terms of convergence speed, from fastest to slowest is as follows: ANJ, APNJ, ADJ, APDJ, BLP.

To establish a context for these methods, consider a standard multinomial logit model:

$$U_{ij}^* = \delta_j + x_{ij}'\beta + \epsilon_{ij} \tag{4.2}$$

$$Y_i = j \quad \text{if} \quad U_{ij}^* \geq \max(U_{ik}^*)\, \forall k \in C, \tag{4.3}$$

for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, J$, where i denotes the decision maker, and j denotes the alternative from the choice set C.

Equation (4.2) models U_{ij}^*, the latent utility for decision maker i from alternative j, with an alternative-specific constant δ_j, vector of observed characteristics, x_{ij}, vector of unknown coefficients, β, and an unobserved disturbance term, ϵ_{ij}. The disturbance terms are distributed i.i.d. type 1 extreme value across decision makers and alternatives. Note that the vector x_{ij} varies across both decision makers and alternatives, but vectors of characteristics that only vary with the decision makers or alternatives can also be used. Also, note that $\delta_1 = 0$ for identification reasons.

A simple method to estimate the parameters of this model, β and $\delta = (\delta_2, \delta_3, \ldots, \delta_J)$, is to maximize the observed log-likelihood $L(\beta, \delta) = \sum_{i=1}^{N} y_{ij} \log(P_{ij})$ with respect to β and δ, where $y_{ij} = 1$ when the i-th decision maker is observed to choose alternative j and 0 otherwise, and

$$P_{ij} = \frac{e^{\delta_j + x'_{ij}\beta}}{\sum_{r=1}^{J} e^{\delta_r + x'_{ir}\beta}}. \tag{4.4}$$

is the logit probability of decision maker i choosing alternative j. When the number of alternatives is large, this method of estimation is difficult. For example, if $J = 1000$, then ML estimation would require a maximization routine over at least 1000 parameters (including the β), which would also require a large gradient vector and Hessian matrix.

Alternatively, we can use the contraction mapping relationship and just search over the space of β. BLP proved that, conditional on β, the vector of alternative-specific constants, δ, is uniquely determined by the known market shares. Therefore, an equivalent way of estimating the unknown parameters in the multinomial logit model is to maximize the log-likelihood $L(\beta, \delta(\beta)) = \sum_{i=1}^{N} y_{ij} \log(P_{ij})$ with respect to β subject to the market share constraints $s_j = N^{-1} \sum_{i=1}^{N} P_{ij}$ for $j = 2, 3, \ldots, J$. This is an equivalent problem, because the market share constraints are the first order conditions of the log-likelihood function with respect to δ for a fixed β. So, this approach is similar concentrating δ out of the likelihood.

In terms of implementation, the optimization algorithm would search over the space of β, but for each value of β, the vector δ is solved for using the market share constraints and the contraction mapping algorithm. The contraction mapping is implemented as the iterative system

$$\delta_j^{(k+1)} \quad = \quad \delta_j^{(k)} + h(\delta^{(k)})f(\delta^{(k)}), \tag{4.5}$$

where the vector-valued function f has elements of the form

$$\delta_j^{(k)} + \log(s_j/N^{-1}\sum_{i=1}^{N} P_{ij}(\delta^{(k)})).$$

The iterative system is iterated until $\parallel \delta^{(k+1)} - \delta^{(k)} \parallel < c$, where c is a convergence criterion that needs to be set to 10^{-14} or smaller following Dube et al. (2011).

The Jacobian matrix for f can be shown to equal

$$f'(\delta) = \begin{pmatrix} 1 - \frac{\sum_{i=1}^{N} P_{i2}^2}{\sum_{i=1}^{N} P_{i2}} & -\frac{\sum_{i=1}^{N} P_{i2}P_{i3}}{\sum_{i=1}^{N} P_{i2}} & \dots & -\frac{\sum_{i=1}^{N} P_{i2}P_{iJ}}{\sum_{i=1}^{N} P_{i2}} \\ -\frac{\sum_{i=1}^{N} P_{i3}P_{i2}}{\sum_{i=1}^{N} P_{i3}} & 1 - \frac{\sum_{i=1}^{N} P_{i3}^2}{\sum_{i=1}^{N} P_{i3}} & \dots & -\frac{\sum_{i=1}^{N} P_{i3}P_{iJ}}{\sum_{i=1}^{N} P_{i3}} \\ \vdots & \dots & \dots & \vdots \\ -\frac{\sum_{i=1}^{N} P_{iJ}P_{i2}}{\sum_{i=1}^{N} P_{iJ}} & -\frac{\sum_{i=1}^{N} P_{iJ}P_{i3}}{\sum_{i=1}^{N} P_{iJ}} & \dots & 1 - \frac{\sum_{i=1}^{N} P_{iJ}^2}{\sum_{i=1}^{N} P_{iJ}} \end{pmatrix}. \tag{4.6}$$

From (4.6), the step sizes $h_{ADJ}(\delta^{(k)})$ and $h_{ANJ}(\delta^{(k)})$ can be constructed easily.

The approximated step sizes $h_{APDJ}(\delta^{(k)})$ and $h_{APNJ}(\delta^{(k)})$ are based on known market shares around the fixed point and the assumption of homogenous decision makers. Specifically, from the market share equations, we know that $s_j = N^{-1}\sum_{i=1}^{N} P_{ij}$ must hold for $j = 2, 3, \dots, J$ around the fixed point, therefore we can approximate all the denominators of the form $\sum_{i=1}^{N} P_{ij}$ from (4.6) with $s_j \times N$. Next, by assuming that each decision maker behaves like the market shares, we can approximate $\sum_{i=1}^{N} P_{ij}P_{ik}$ from (4.6) with $s_j \times s_k \times N$.

With these assumptions, $f'(\delta)$ can be approximated as

$$a(MS) = \begin{pmatrix} 1-s_2 & -s_3 & \dots & -s_J \\ -s_2 & 1-s_3 & \dots & -s_J \\ \vdots & \dots & \dots & \vdots \\ -s_2 & -s_3 & \dots & 1-s_J \end{pmatrix}. \tag{4.7}$$

This approximated matrix is used to construct $h_{APDJ}(\delta^{(k)})$ and $h_{APNJ}(\delta^{(k)})$. The approximations are better when the iterative system is close to the fixed point and when the decision makers behave like the market shares. Even if these assumptions do not hold exactly, the resulting iterative systems should still be faster than the BLP system. Another feature of this approximation is that it does not depend on δ. Unlike the analytic modifications, it is not necessary to recalculate these step sizes at each δ or in each iteration of the system. This saves a lot of time because calculations of the partial derivatives and inversions matrices are avoided. The biggest improvements occur when J is large. Also, from the simulation study, this matrix is numerically more stable than the analytic Jacobian.

4.4 Simulation study

I perform a Monte Carlo simulation study to analyze the performance of the various algorithms. Specifically, I generate 500 sets of data for a conditional logit model, and for each set of data, the δ vector from the data generating process (conditional on the true β) is recovered using the five algorithms (BLP and the four proposed algorithms).

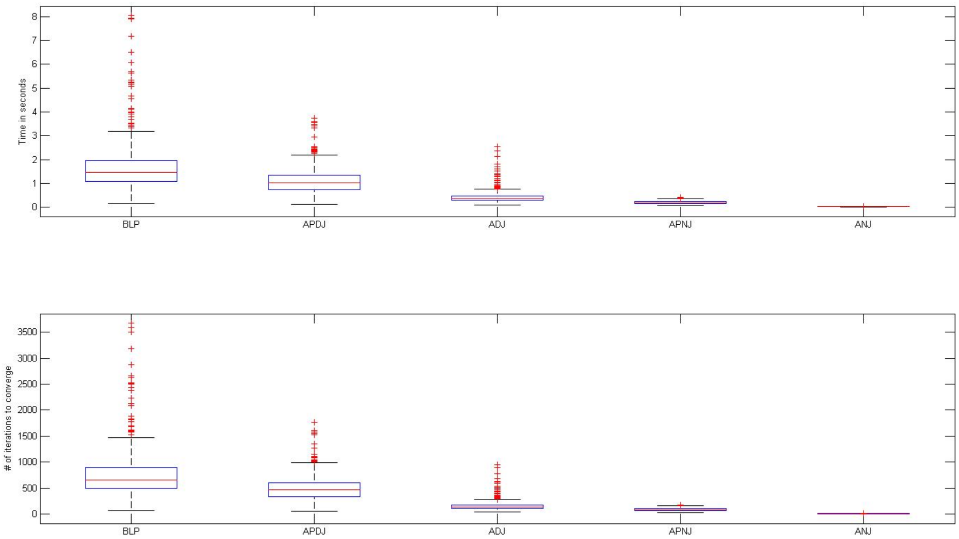

Figure 4.1: Boxplots for computational time and number of iterations until convergence for 500 sets of data. The convergence criterion is set to 10^{-14}.

I use $N = 5000$ decision makers and $J = 6$ alternatives. For identification reasons, δ_1 is set to 0, which means only $\delta = (\delta_2, \delta_3, \ldots, \delta_6)$ needs to be recovered. For each set of data, the elements of the "true" δ and covariate vectors x_{ij} are drawn randomly from a Normal distribution with mean 0 and standard deviation 2. The starting vector for each algorithm is set to a vector of zeros. The convergence criterion is set to 10^{-14}.

Figure 4.1 contains boxplots of the computational time and number of iterations required until convergence over 500 sets of data. These plots suggest that the BLP contraction mapping is quite slow compared to the proposed methods. The median number of iterations until convergence for BLP is 653, while the medians for APDJ, ADJ, APNJ, ANJ are 469, 139, 84, and 8, respectively. The analytic Newton (ANJ) step size reduces the median number of iterations required until converge by approximately 80 times relative to BLP. Similarly, the approximated Newton (APNJ) step size needs approximately 8 times fewer iterations to converge.

The figure also reveals that the number of iterations it takes for the BLP method to converge has large variability while the ANJ and APNJ methods do not. In this study, the BLP method can at times require up to 3600 iterations to converge while the ANJ and APNJ methods only need 196 and 14 iterations, respectively. At the extremes, these two proposed methods offer approximately 20-fold and 260-fold improvements over BLP.

Therefore, I recommend using the analytic Newton step size (ANJ) whenever it is possible. But, if it takes too much time to compute, then the approximated Newton step size (APNJ) offers somewhat comparable performance. Otherwise, the two remaining approaches, APDJ and ADJ, also provide moderate improvements over the BLP contraction mapping. See Table 4.2 for their relative merits.

4.5 Concluding remarks and future research

The four proposed methods are all faster to converge than the BLP contraction mapping. I recommend the analytic Newton (ANJ) step size whenever possible, but the approximated Newton (APNJ) step size offers good performance as well. If local convergence is a concern, they can be combined with the BLP step size matrix. For example, use the BLP step size matrix in the initial iterations of the system, then switch over to one of the proposed step sizes in the later iterations. This approach uses the global convergence feature of the BLP method to bring the initial iterates closer to the fixed point first and then capitalizes on the increased speed of the proposed methods after that. This hybrid method works extremely well and is recommended when there are suspected convergence issues.

These new methods are promising as they are easy to code and result in less variability in terms of the number of iterations required to converge. Future research direction includes extending this analysis to mixed logits and comparing the results to the constrained maximum likelihood approach of Dube et al. (2011).

Bibliography

Albert, J. H. and S. Chib (1993): "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

——— (2001): "Sequential Ordinal Modeling with Applications to Survival Data," *Biometrics*, 57, 829–836.

Amemiya, T. (1984): "Tobit models: A survey," *Journal of Econometrics*, 24, 3–61.

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

Berry, S. T. (1994): "Estimating Discrete-Choice Models of Product Differentiation," *RAND Journal of Economics*, 25, 242–262.

Bhat, C. and N. Eluru (2009): "A copula-based approach to accommodate residential self-selection effects in travel behavior modeling," *Transportation Research Part B*, 43, 749–765.

Bőrsch-Supan, A. and V. A. Hajivassiliou (1993): "Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models," *Journal of Econometrics*, 58, 347 – 368.

Boyes, W. J., D. L. Hoffman, and S. A. Low (1989): "An econometric analysis of the bank credit scoring problem," *Journal of Econometrics*, 40, 3 – 14.

Brownstone, D. and A. Fang (2009): "A Vehicle Ownership and Utilization Choice Model With Endogenous Residential Density," Working papers, University of California, Irvine.

Brownstone, D. and T. F. Golob (2009): "The impact of residential density on vehicle usage and energy consumption," *Journal of Urban Economics*, 65, 91–98.

Cervero, R. and K. Kockelman (1997): "Travel demand and the 3Ds: Density, diversity, and design," *Transportation Research Part D: Transport and Environment*, 2, 199 – 219.

Chan, J. C.-C. and I. Jeliazkov (2009): "MCMC Estimation of Restricted Covariance Matrices," *Journal of Computational and Graphical Statistics*, 18, 457–480.

Chen, M. and D. Dey (2000): *Generalized Linear Models: A Bayesian Perspective*, CRC Press, chap. Bayesian Analysis for Correlated Ordinal Data Models.

Chib, S. (2001): "Markov chain Monte Carlo methods: computation and inference," in *Handbook of Econometrics*, ed. by J. Heckman and E. Leamer, Elsevier, vol. 5 of *Handbook of Econometrics*, chap. 57, 3569–3649.

——— (2007): "Analysis of treatment response data without the joint distribution of potential outcomes," *Journal of Econometrics*, 140, 401–412.

Chib, S. and E. Greenberg (1995): "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 49, 327–335.

——— (1998): "Analysis of Multivariate Probit Models," *Biometrika*, 85, pp. 347–361.

Chib, S., E. Greenberg, and I. Jeliazkov (2009): "Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection," *Journal of Computational and Graphical Statistics*, 18, 321–348.

Dube, J.-P., J. T. Fox, and C.-L. Su (2011): "Improving the Numerical Performance of BLP Static and Dynamic Discrete Choice Random Coefficients Demand Estimation," *Econometrica*, Forthcoming.

Dunphy, R. and K. Fisher (1996): "Transportation, Congestion, and Density: New Insights," *Transportation Research Record: Journal of the Transportation Research Board*, 1552, 89–96.

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, 736–758.

Fréchet, M. (1951): "Sur les Tableaux de Correlation Dont les Marges son Donnees," *Annals of the University of Lyon*, Sec. A, 53–77.

Genius, M. and E. Strazzera (2008): "Applying the copula approach to sample selection modelling," *Applied Economics*, 40, 1443–1455.

Geweke, J. (1991): "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints," *Computing Science and Statistics*, 571–578.

Greenberg, E. (2007): *Introduction to Bayesian Econometrics*, Cambridge University Press.

Greene, W. H. (2008): *Econometric Analysis*, Prentice Hall.

Gronau, R. (1973): “The Effect of Children on the Housewife’s Value of Time,” *The Journal of Political Economy*, 81, S168–S199.

Hajivassiliou, V. A. and D. L. McFadden (1998): “The Method of Simulated Scores for the Estimation of LDV Models,” *Econometrica*, 66, pp. 863–896.

Heckman, J. (1979a): “Sample Selection Bias as a Specification Error,” *Econometrica*, 47, 153–61.

Heckman, J. J. (1976): “The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models,” in *Annals of Economic and Social Measurement, Volume 5, number 4*, National Bureau of Economic Research, Inc, NBER Chapters, 120–137.

——— (1979b): “Sample Selection Bias as a Specification Error,” *Econometrica*, 47, 153–61.

——— (1990): “Varieties of Selection Bias,” *The American Economic Review*, 80, 313–318.

Hoeffding, W. (1940): “Masstabinvariente Korrelationstheorie,” Schriften des Mathematischen Instituts und des Instituts fur Angewandt Mathematik der Universitat Berlin.

Jeliazkov, I., J. Graves, and M. Kutzbach (2008): “Fitting and Comparison of Models for Multivariate Ordinal Outcomes,” in *Advances in Econometrics: Bayesian Econometrics*.

Jeliazkov, I. and E. Lee (2010): “MCMC perspectives on simulated likelihood estimation,” *Advances in Econometrics: Maximum Simulated Likelihood Methods and Applications*, 3–39.

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

Judd, K. L. and B. S. Skrainka (2011): “High performance quadrature rules: how numerical integration affects a popular model of product differentiation,” CeMMAP working papers CWP03/11, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Keane, M. P. (1994): “A Computationally Practical Simulation Estimator for Panel Data,” *Econometrica*, 62, 95–116.

Knittel, C. R. and K. Metaxoglou (2012): “Estimation of Random Coeffcient Demand Models: Two Empiricists’ Perspective,” Working paper, MIT.

Koop, G., D. Poirier, and J. Tobias (2007): *Bayesian Econometric Methods*, Cambridge University Press.

Lee, E. H. (2010): "Copula, Simulated Likelihood, and VAR Estimation based on MCMC," Ph.D. thesis, University of California, Irvine.

Lee, L.-F. (1983): "Generalized Econometric Models with Selectivity," *Econometrica*, 51, 507–12.

Li, P. (2011): "Estimation of sample selection models with two selection mechanisms," *Computational Statistics and Data Analysis*, 55, 1099–1108.

Liu, J. S. (1994): "The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem," *Journal of the American Statistical Association*, 89, 958–966.

Manski, C. F. (1989): "Anatomy of the Selection Problem," *The Journal of Human Resources*, 24, 343–360.

McFadden, D. (1973): "Conditional logit analysis of qualitative choice behavior," *Frontiers in Econometrics*, 105–142.

Musalem, A., E. T. Bradlow, and J. S. Raju (2009): "Bayesian estimation of random-coefficients choice models using aggregate data," *Journal of Applied Econometrics*, 24, 490–516.

Nandram, B. and M.-H. Chen (1996): "Reparameterizing the generalized linear model to accelerate gibbs sampler convergence," *Journal of Statistical Computation and Simulation*, 54, 129–144.

Nelsen, R. B. (1998): *An Introduction to Copulas (Lecture Notes in Statistics)*, Springer, 1 ed.

Nevo, A. (2000): "A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand," *Journal of Economics and Management Strategy*, 9, 513–548.

Newey, W. K., J. L. Powell, and J. R. Walker (1990): "Semiparametric Estimation of Selection Models: Some Empirical Results," *The American Economic Review*, 80, 324–328.

Olver, P. J. (2006): "Nonlinear Systems," Lecture notes, School of Mathematics, University of Minnesota.

Park, S. and S. Gupta (2009): "Simulated Maximum Likelihood Estimator for the Random Coefficient Logit Model Using Aggregate Data," *Journal of Marketing Research*, XLVI, 531–542.

——— (2011): "Comparison of SML and GMM estimators for the random coefficient logit model using aggregate data," *Empirical Economics*, 1–20, 10.1007/s00181-011-0519-3.

Pitt, M., D. Chan, and R. Kohn (2006): "Efficient Bayesian inference for Gaussian copula regression models," *Biometrika*, 93, 537–554.

Poirier, D. J. (1980): "Partial observability in bivariate probit models," *Journal of Econometrics*, 12, 209–217.

Prieger, J. (2000): "A Generalized Parametric Selection Model for Non-normal Data," Working Papers 00-9, University of California at Davis, Department of Economics.

Puhani, P. A. (2000): "The Heckman Correction for Sample Selection and Its Critique," *Journal of Economic Surveys*, 14, 53–68.

Rothenberg, T. J. (1971): "Identification in Parametric Models," *Econometrica*, 39, pp. 577–591.

Shonkwiler, J. S. and S. T. Yen (1999): "Two-Step Estimation of a Censored System of Equations," *American Journal of Agricultural Economics*, 81, 972–982.

Sklar, A. (1959): "Fonctions de repartition a n dimensions et leurs marges," *Publications de l'Institut de Statistique de L'Universit de Paris*, 8, 229–231.

——— (1973): "Random variables, joint distributions, and copulas," *Kybernetica*, 9, 449–460.

Skrainka, B. S. (2012): "A Large Scale Study of the Small Sample Performance of Random Coefficient Models of Demand," Working paper.

Smith, M. D. (2003): "Modelling sample selection using Archimedean copulas," *Econometrics Journal*, 6, 99–123.

Song, P. X.-K. (2000): "Multivariate Dispersion Models Generated from Gaussian Copula," *Scandinavian Journal of Statistics*, 27, 305–320.

Tanner, M. A. and W. H. Wong (1987): "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540.

Terza, J. V. (1998): "Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects," *Journal of Econometrics*, 84, 129 – 154.

Tierney, L. (1994): "Markov chains for exploring posterior distributions," *Annals of Statistics*, 22, 1701–1762.

van Hasselt, M. (2009): "Bayesian Inference in a Sample Selection Model," Working papers, The University of Western Ontario.

Vella, F. (1998): "Estimating Models with Sample Selection Bias: A Survey," *The Journal of Human Resources*, 33, pp. 127–169.

Wooldridge, J. M. (1998): "Selection Corrections with a Censored Selection Variable," Mimeo, Michigan State University Department of Economics.

——— (2002): *Econometric Analysis of Cross Section and Panel Data*, MIT press.

Yang, S., Y. Chen, and G. M. Allenby (2003): "Bayesian Analysis of Simultaneous Demand and Supply," *Quantitative Marketing and Economics*, 1, 251–275, 10.1023/B:QMEC.0000003327.55605.26.

Yen, S. T. (2005): "A Multivariate Sample-Selection Model: Estimating Cigarette and Alcohol Demands with Zero Observations," *American Journal of Agricultural Economics*, 87, 453–466.

Zimmer, D. M. and P. K. Trivedi (2005): *Copula Modeling: An Introduction for Practitioners*, Foundations and Trends in Econometrics.

——— (2006): "Using Trivariate Copulas to Model Sample Selection and Treatment Effects: Application to Family Health Care Demand," *Journal of Business & Economic Statistics*, 24, 63–76.