Discrete Choice Models - Estimation of Passenger Traffic

Sørensen, Majken Vildrik

Publication date:
2003

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Discrete Choice Models
Estimation of Passenger Traffic
Ph.D. thesis

Majken Vildrik Sørensen
Centre for Traffic and Transport
Technical University of Denmark,
DK-2800 Kgs. Lyngby, Denmark

May, 2003
Preface

This thesis concludes the Ph.D. study on traffic model undertaken by Majken Vildrik Sørensen between 2000-2003. The objective of the study was to improve models to estimate mode choice for passenger traffic by improved understanding of which aspects determines the choice of transport mode for an individual. Improvement of the utility function describing the alternatives, functional forms to model the choice of individuals and estimation procedures has been included in the study.

The Ph.D study was undertaken at Centre for Traffic and Transport at the Danish Technical University in the period 2000-2003. Otto A. Nielsen, Center for Traffic and Transport, Danish Technical University, supervised the study and Poul Thyregod, Institute of Mathematical Modelling, Danish Technical University, were additional supervisor within areas of statistics.

The study was funded by a conglomerate of companies; DSB Passagertog (the National Rail Operator) 50 %, Banestyrelsen (the National Rail Agency) 25 % and Atkins Danmark (formerly Banestyrelsen rådgivning) 25 %.

Two different methods have been proposed which are related in structure but applied to different problems. The first method (further described in section 5) concerns the data segmentation often applied in practical modelling to ensure a better fit of model to the data. A new method for data segmentation based on respondents actual behaviour was proposed (Sørensen, 2001). The second method (further described in section 6) concerns the issue of individuals having different preferences which results in a Mixed Logit formulation. The proposed method seek to assess the empirical distribution of preferences from a sample (population), described in (Sørensen and Nielsen, 2001, 2002) (joint work with O. A. Nielsen) and extended to alternative specifications of the utility function in (Sørensen, 2002).

During the study, six papers have been written and were either presented on conferences or submitted to a journal. The papers vary in theme and are centered around two different traffic models covering eastern Denmark and southern Sweden (the Copenhagen Ringsted Model and the Øresund Model). The Copenhagen Ringsted Model is a large scale, stochastic traffic model, was reviewed
with regards to statistical performance in particular convergence, stability and reproductiveness (Sørensen et al., 2001) (joint work with O. A. Nielsen and D. Filges). The Øresund Model was reviewed in the light of a before and after study of the model, (Sørensen et al., 2002) (joint work with O. A. Nielsen and J. Schauby).

Acknowledgements

Atkins Scandinavia Transportation, the National Rail Agency (Banestyrelsen) and the Danish Rail Operator (DSB) are thanked for funding the Ph.D.-study. Otto A. Nielsen, Center for Traffic and Transport at Danish Technical University for being my supervisor and Poul Thyregod, Institute of Mathematical Modelling, DTU for being additional supervisor. Atkins Scandinavia Transportation is thanked for providing data and existing code for the work. Andrew Daly, Rand Europe is thanked for valuable discussions and Moshe Ben-Akiva, MIT, Boston for letting me visit the ITS Lab at MIT. An finally, Øystein Leonardsen is thanked for putting a lot of effort into getting the project started, though various political agendas almost prevented him from succeeding.

Copenhagen, May 2003.

Majken Vildrik Sørensen
Abstract

Discrete Choice Models. Estimation of Passenger Traffic

This thesis gives an overview of what has been done in the research area of passenger transport modelling, with a focus on the model type in the core of a model complex. After a formulation of the choice problem (choice probability, the set alternatives), a method for estimation and requirements for data, a literature review follows. Models applied for estimation of discrete choice models are described by properties and limitations, and relations between these are established. Model types are grouped into three classes, Hybrid choice models, Tree models and Latent class models. Relations between model, data and estimation are described, with a focus of possibilities/limitations of different techniques.

Two special issues of modelling are addressed in further detail, namely data segmentation and estimation of Mixed Logit models. Both issues are concerned with whether individuals can be assumed 'homogeneous', that is, can be described by the same model (fixed coefficients). First, a new method for data segmentation is proposed, which segments data by individual preferences. Segmentation by individual preferences will diminish the severeness of the assumed homogeneity of individuals (assumed for estimation of choice models). For application of the method an algorithm is provided with a case.

Also for the second issue, estimation of Mixed Logit models, a method was proposed. The most commonly used approach to estimate Mixed Logit models, is to employ the Maximum Simulated Likelihood estimation (MSL), which simultaneously finds optimal coefficients values (utility elements) and parameter values (distributed terms) in the utility function. The shape of the distributed terms is specified prior to the estimation; hence, the validity is not tested during the estimation. The proposed method, assesses the shape of the distribution from data, by means of repetitive model estimation. In particular, one model was estimated for each sub-sample of data. The shape of distributions is assessed from between model comparisons. This is not to be regarded as an alternative to MSL estimation, rather as a complimentary test to assess the shape of distribution prior to a MSL estimation.

The method is tested on synthetic data where different shapes of distribution are
assumed for distributed terms. Differences in choice of alternative are asserted for different distributions and MNL estimation is seen to have some difficulty in explaining the choice of alternative. For the MSL estimation, some problems with the traditional determination of which shape of distribution to employ, are demonstrated. Further, all tested distributed terms (different shape of distribution) were significant; neither of these distributions were superior to the others. The proposed method for determination of shape of distribution, was able to recover the 'correct' shape of distributions and to pinpoint which term had the highest level of variance and therefore seems very promising. Following the method was applied to 'real' data, where distribution of coefficients were found. All the shapes of distributions found, complied with sound knowledge in terms of which should be uni-modal, sign specific and/or skewed distributions.

The thesis concludes by summing up the results and recapturing the areas where further research is needed.
Diskrete valgmodeller. Estimation af passagertrafik


Metoden er testet på syntetisk data hvor forskellige former er antaget for de
fordelte led. For forskellige datasæt er der forskelle mellem hvilket alternativ der er valgt, ligeledes ses der problemer med estimation af MNL modellen. For MSL estimationen er der konstateret problemer med den måde hvorpå formen for fordelte led traditionelt er blevet bestemt. Endvidere er alle de testede statistiske former for fordelingen signifikante; mens ingen af disse er påviselig bedre end øvrige former. Den fremsatte metode til bestemmelse af form for fordelingen, var i stand til at korrekt bestemme formen for fordelingen, og ydermere korrekt bestemme hvilket fordelt led, der havde højst/lavest varians. Methoden ser derfor særdeles lovende ud. Efterfølgende er metoden blevet anvendt på empirisk data, hvor fordeling blev påvist (fordelte koefficienter). De påviste former var alle i overensstemmelse med 'sund fornuft' mht. uni-modalitet, fortægnspecifikke fordelingen og/eller skæve fordelinger.

Afhandlingen slutter med en opsummering af resultater samt en opsummering af områder hvor der er behov for yderligere forskning.
# Contents

1 Introduction .................................................. 1  
  1.1 Purpose .................................................. 1  
  1.2 Outline .................................................. 2  
  1.3 General notation ......................................... 3  

2 The behaviour of choice .................................... 5  
  2.1 Choice strategies .......................................... 5  
  2.2 Model estimation ......................................... 7  
    2.2.1 Maximum Likelihood ................................... 7  
    2.2.2 Properties of the estimator .......................... 10  
  2.3 Choice context ........................................... 11  
    2.3.1 Extending complexity of models ..................... 13  
    2.3.2 Identification of models ............................ 14  

3 Existing models ............................................. 17  
  3.1 Background ................................................ 17  
  3.2 Models .................................................... 18  
  3.3 The Logit model .......................................... 19  
  3.4 The GEV family .......................................... 20  
  3.5 Models with correlated alternatives .................... 21  
    3.5.1 The Nested Logit Model .............................. 22  
    3.5.2 The Joint Logit model .............................. 24  
    3.5.3 The Generalised Nested Logit ....................... 24  
    3.5.4 The Ordered Generalised Extreme Value model .... 26  
    3.5.5 Ordered models ...................................... 27  
    3.5.6 The Covariance Heterogeneity Nested Logit model .. 27  
    3.5.7 The Recursive Nested Extreme Value model ....... 28  
    3.5.8 The C-Logit model .................................. 29  
    3.5.9 The Probit model ................................... 29  
    3.5.10 Latent Preference models .......................... 30  
  3.6 Models allowing for non-deterministic behaviour ....... 32  
    3.6.1 The Random Coefficients Logit model ............... 32  
    3.6.2 The Mixed Logit model ............................ 34  
  3.7 Models allowing for heteroscedasticity ................ 34
## CONTENTS

3.7.1 The Heteroscedastic Extreme Value model .......................... 35
3.7.2 The Heterogeneous Conditional Logit model ...................... 36
3.8 Elasticities ...................................................................... 36
3.9 Discussion ...................................................................... 36

4 Data and estimation .......................................................... 41
4.1 Data ............................................................................ 41
4.2 Data quality ................................................................... 44
  4.2.1 Data types .............................................................. 46
4.3 Data Segmentation ............................................................. 48
  4.3.1 Mathematically ......................................................... 49
  4.3.2 'Hierarchical' segments ............................................... 51
4.4 Sampling techniques ......................................................... 52
  4.4.1 Random sampling ..................................................... 53
  4.4.2 Cluster analysis ......................................................... 54
4.5 Closed form estimation ....................................................... 57
4.6 Simulation based techniques ................................................. 59
  4.6.1 Maximum simulated likelihood .................................... 59
  4.6.2 Quadrature .............................................................. 60
  4.6.3 Random numbers ..................................................... 62
4.7 Sample based techniques ...................................................... 62
  4.7.1 Jackknife .................................................................. 62
  4.7.2 Bootstrapping .......................................................... 63
4.8 Forecast ........................................................................ 64
  4.8.1 Segmented model ....................................................... 66
4.9 Summary ...................................................................... 67

5 Data segmentation by preferences (SEP Method) ......................... 71
5.1 Introduction .................................................................... 71
  5.1.1 Background .............................................................. 72
5.2 SEP method for data segmentation .................................... 74
  5.2.1 Purpose of the SEP method ....................................... 74
  5.2.2 Motivation for the SEP method ................................... 74
  5.2.3 Existence ............................................................... 77
5.3 Algorithm .................................................................... 78
  5.3.1 Data grouping .......................................................... 80
  5.3.2 Identification of segments ......................................... 82
  5.3.3 Hierarchical segments ............................................... 85
5.4 Discussion .................................................................... 85
5.5 Empirical evidence ............................................................ 88
  5.5.1 Summary of empirical evidence ................................. 90
## CONTENTS

6 Assessment of shape of distributions (SODA Method) 91  
6.1 Introduction and background .......................... 91  
6.2 SODA method for assessment of shape of distribution 92  
6.2.1 Motivation for SODA method ...................... 94  
6.3 Algorithm ........................................... 94  
6.3.1 Data grouping ..................................... 97  
6.3.2 Shape of distribution ............................. 98  
6.3.3 Moments and covariance .......................... 101  
6.4 Discussion .......................................... 102  

7 Application 107  
7.1 Synthetic data ...................................... 107  
7.1.1 Description ...................................... 109  
7.1.2 Comparisons of synthetic data sets ............... 112  
7.1.3 MNL models of synthetic data sets ............... 112  
7.1.4 MSL models of synthetic data sets ............... 117  
7.1.5 Results by SODA method ........................... 120  
7.1.6 50-500 groups, 1 and 3 coefficients .......... 134  
7.1.7 Comparisons of synthetic data .................. 141  
7.1.8 Conclusion ........................................ 147  
7.2 The Copenhagen-Ringsted case ....................... 149  
7.2.1 Data description ................................... 149  
7.2.2 Test 1, 100 groups ................................ 150  
7.2.3 Test 2, 50 groups, 10 runs ....................... 158  
7.2.4 Test 3, 25 groups, 10 runs ....................... 158  
7.2.5 All tests .......................................... 158  
7.2.6 Stability of results ............................... 159  
7.2.7 Conclusions of CRM data ........................ 159  

8 Conclusion 165  

A Statistics 171  
A.1 Moments ............................................. 171  
A.1.1 Moments calculated from sub-samples .......... 172  
A.1.2 Correlation ....................................... 173  
A.2 Gumbel (or Extreme Value, Type I) distribution 173  
A.3 Log-normal distribution ............................. 175  
A.4 Test for distribution ................................ 176  

B Alternative segmentation of CRM data 177  
B.1 Results ............................................. 177  
B.2 Summary ............................................ 180  

C Figures for distributed coefficients 183
Chapter 1

Introduction

1.1 Purpose

The purpose of the Ph.D. study, was to improve models ability to estimate traffic demand through improved understanding of which factors contributes to explain the choice between modes. For this, a statistical model should be formulated (based on data) and it should be evaluated whether it can explicitly handle additional statistical characteristics like e.g. 'distributed coefficients', correlation between variables. New model variants should also, be included.

This task was structured as three different objectives. First, to get an overview of the current theory and methodologies for modelling transport. Second, to pinpoint areas of special interest with room for improvement, and finally, to attempt to come up with specific ideas for improvement.

The intention with the literature review undertaken, was to collate the numerous models applied for estimation of traffic models, or more specifically, discrete choice models for transport mode choice for passengers. These models are not solely applied for transport mode choice; further these are employed for route choice (assignment), choice of destination, location or in any other context where exactly one alternative is chosen from a set of alternatives (e.g. marketing).

Through improved insight into the theory behind the models, shortfalls with methods applied, were discovered. And fortunately, during the study solutions came up for two of these. Hereby, the three objectives outlined above have been covered, and this is demonstrated by this Ph.D. thesis.
1.2 Outline

Below is a brief outline of the structure of this thesis.

Chapter 1
Introduce and outline the thesis.

Chapter 2
Describes how the choice process is handled, by the individual and by researchers in the modelling step. The relation between choice context (alternatives and choice strategies) and the model estimation is described in graphics.

Chapter 3
This chapter is a literature review of model variants, where model names and abbreviations are linked. Models are described by assumptions and limitations and a ‘tree’ of the variants are given to give an overview of relations between models.

Chapter 4
Data and estimation techniques (closed form, simulation based and sample based) are briefly described, as any model is dependent on the data it is estimated on and the accuracy of a model estimation is dependent on the estimation technique employed. The chapter concludes with model results impact on forecast.

Chapter 5
In this chapter a new method (SEP) for data segmentation is proposed. The method (heuristic) is related to ‘traditional’ segmentation, though the key is segmentation by individual preferences. An algorithm is formulated and test results are discussed.

Chapter 6
A crucial assumptions for estimation of Mixed Logit Models (distribution of distributed terms) is discussed and a method (SODA) for empirical determination the distribution prior to the estimation is proposed. An algorithm is formulated.

Chapter 7
The SODA method proposed in the previous chapter is applied to first a number of synthetic data sets, to determine the SODA methods capability to determine the distribution from data. Subsequently, the method is applied to ‘real’ data.

Chapter 8
A conclusion of the study, and of the thesis.

Appendix A
Statistical prerequisites for calculation and relevant statistical distributions.
1.3 General notation

The term *coefficients* is used to refer to the term multiplied to variables in the utility function (the $\beta$ in $U = \sum_k \beta_k X_k$); the term *parameter* refers to the description of a statistical distribution, i.e. $\alpha, \sigma$ in $N(\alpha, \sigma^2)$.

\begin{align*}
n & \quad \text{Individuals} \ (n = 1, \ldots, N) \\
i & \quad \text{Alternatives} \\
g & \quad \text{Groups, sub-samples of data} \ (g = 1, \ldots, G) \\
m & \quad \text{Segments} \ (m = 1, \ldots, M) \\
C & \quad \text{The set of all alternatives} \\
X_k & \quad \text{Explanatory variables, covariates, attributes} \ (k = 1, \ldots, K) \\
X^m & \quad \text{Subset of data, related to segment} \ m \\
U_i & \quad \text{The utility for alternative} \ i \\
V_i & \quad \text{The deterministic part of the utility for alternative} \ i \\
\epsilon_i & \quad \text{The stochastic part of the utility for alternative} \ i \\
P_i & \quad \text{The probability for choosing alternative} \ i, \ P_i = P(U_i \geq U_j, j \neq i) \\
\beta_k & \quad \text{Coefficients} \ (\text{fixed}) \ in \ the \ utility \ function \ (k = 1, \ldots, K) \\
B & \quad \text{A matrix, with} \ \beta_k \ \text{on the diagonal} \\
\xi_k & \quad \text{Stochastic part of a coefficients in the utility function} \ (k = 1, \ldots, K) \\
\Xi & \quad \text{A matrix, with} \ \xi_k \ \text{on the diagonal} \\
\ell & \quad \text{The likelihood} \\
\mathcal{L} & \quad \text{The log-likelihood} \\
\mu & \quad \text{Logit scale parameter} \\
E(X) & \quad \text{The mean value of} \ X \\
Var(X) & \quad \text{The variance of} \ X \\
m_t(X) & \quad \text{The} \ t \ (\text{central}) \ \text{moment of variable} \ X \\
\hat{a} & \quad \text{The} \ '\text{hat}^{' \ \text{denotes estimator of} \ a} \\
N & \quad \text{Normal distribution} \\
LN & \quad \text{Log-Normal distribution} \\
IN & \quad \text{Inverse Normal distribution} \\
U & \quad \text{Uniform distribution}
\end{align*}
Chapter 2

The behaviour of choice

When a person decides to choose one alternative before a number of other alternatives this may or may not, be based on carefully considerations of each of the alternatives with respect to the situation the person is in. It may be based solely on an instant idea of taking the train simply because the person feels like it, or going by car as there may turn up a need for having it. Whatever the 'motivation' for the action, little information is left for the observer. If the same person was studied for a period of time it may even seem as if the person was being inconsistent with him/herself, as different choices may have arise from what seemed to be identical situations. Needless to say, different person are often choosing differently from what seems to be identical situations.

The key into being able to describe the choice of an individual is to understand the 'motivation' behind an action, even if there is no genuine 'motivation'. The following step is to understand whether and how this 'motivation' may differ from one context to another, from time to another etc., and hereafter, to get an understanding of this for the population of interest. It may appear simple, but in practice it may be at the best, close to impossible.

Inferring about the future actions of an individual is inferring about the changes in the 'motivations' of the individual's actions. Inferring about a group of individual future actions generally involves a general description of all the individuals 'motivations' (leaving out some details) as the idea of describing a group by aggregation is to simplify description, and thereby reduce the workload.

2.1 Choice strategies

Modelling of traffic is generally based on empirical data consisting of socio-economic variables, variables describing the alternatives (attributes) and information of chosen alternative - but no (direct) information on why this alternative was chosen among a (known) set of alternatives. Demand models seek to uncover
CHAPTER 2. THE BEHAVIOUR OF CHOICE

the interaction/ correspondence between the descriptive variables (attributes) and the observed (realised) behaviour, i.e. the alternative chosen. To put it in other words, the individuals underlying choice strategy or 'motivation' for the observed choice of alternative is unknown to the modeller - even the individual itself may not be aware of it. Likewise, there is a lack of information of whether all individuals conform to the same choice strategy, whether the same individual conforms to the same strategy at different periods of times or even for different purposes. An example of some different choice strategies follows:

1. Car is always chosen
2. The alternative characterised by the lowest total travel time (door-to-door) is chosen
3. Generally bike - but if it is raining, then bus is chosen
4. The alternative characterised by the highest utility is chosen

Strategy 1, is an example of a dominating strategy, also referred to as the lexicographic rule, while strategy 2 has a dominating characteristic. Likewise has the third strategy a dominating characteristic but normally this information would not be available in data applied for traffic analysis as information on weather conditions is seldom included. Hence, the fourth strategy would normally be assumed to be descriptive for the choices of alternative undertaken by the individual. For the remainder of this thesis this approach is applied.

Put in more technical terms modelling a choice situation involves description of the choice context which includes all available alternatives, and the subset of by the individual considered alternatives, full description of aspects of alternatives considered irrespective of whether the consideration is deliberate or not, time of choice and circumstances of choice (travelling alone or in a group, repetitive choice situation e.g. to/from work, exterior conditions e.g. weather, etc.). Given the above, modelling of the choice situation can be undertaken. Unfortunately, this information is generally not available wherefore simplifying assumptions have to be made in order to model the choice of a population.

Models that aim at describing individual behavior by a combination of utility theory and statistical tools are generally referred to as econometric models. These models are closely related to specification of the utility function (which variables are included in a linear or non-linear function), and then in turn, to the data the model coefficients are estimated from. A framework for choice modelling (choice, alternatives, estimation) is formulated in the following. Then follows a brief discussion of issues of identifiability. Related to specification of the utility is the topic of data segmentation, where separate models often are constructed for e.g. different travel purposes to reflect differences in behaviour for different types of travellers/trips.
2.2 Model estimation

Having defined the framework for modelling of choice a literature review of models proposed through times follows in chapter 3. Estimation methods and data are the topics of chapter 4. Often separate models are formed for different 'segments' of users in order to improve model fit. An alternative to trivial segmentation (by e.g. purpose) is proposed in chapter 5. Estimation methods for stochastic models are another important issue of determination of the 'best model' for a given data set. The frequently used tool of simulation-assisted likelihood analysis is discussed and a tool for improvement of the method is suggested in chapter 6. Applications of the proposed methods follows in chapter 7 and finally, conclusions are in chapter 8.

2.2 Model estimation

Before heading into complex discussions of distributions of several error terms or tree structures with extensive cross-nesting, a basic description of the utility and estimation procedure is given. Specify a utility function as

\[ U_{ji} = V_{ji} + \nu_j \]  

(2.1)

where \( V_{ji} \) respective \( \nu_j \) represents the explained (deterministic) respective the unexplained (stochastic) variation in the utility function for choice of alternative \( i \) by individual \( j \). The explained variation is a function of some explanatory variables \( X \), weighted by coefficients \( \beta \), that is \( V_{ji} = f(\beta; X) \), and is often formulated as a linear relation, i.e. \( V_{ji} = \sum_k \beta_k X_{jik} \) where each of the \( k = 1, \ldots, K \) attributes \( X_k \) are converted into an arbitrary measure utility. Alternatively, non-linear formulations may be used e.g. raised to a power, Box-Cox transformations\(^1\) or Box-Tukey\(^2\). Descriptions of these transformers are found in e.g. Ben-Akiva and Lerman (1985).

This framework is (usually) operationalised by assuming that the unexplained part of the variation is described by some well-known distribution (Extreme Value, type I or Normal). Estimation is performed by means of maximum likelihood, where the estimates of the coefficients are conditional on the specification of the utility function (which attributes are included) as well as the \textit{a priori} (assumed) distribution of the \( \nu \).

2.2.1 Maximum Likelihood

For estimation of discrete choice models, the method of maximum likelihood is applied. For the sake of illustrating the method, the functional form of the choice model (the probability for choosing alternative \( i \) from a set \( C \)), is described by \( P_i = P_i(U_i) \). The likelihood function can be specified as a function of this

---

\(^1\)The Box-Cox transform is defined as \( y(\lambda) = (y^\lambda - 1)/\lambda \) for \( \lambda \neq 0 \); \( \log y \) otherwise. The transformation is continuous and well-defined for \( y > 0 \).

\(^2\)The Box-Tukey transform is defined as \( ((y + \alpha)^\lambda - 1)/\lambda \).
choice probability, which again is dependent on the explanatory variables and some coefficients.

\[
\ell = \ell(\beta; X) = \prod_n \prod_i (P_i)^{g_{in}} \tag{2.2}
\]

where \( g_{in} \) is 1 if alternative \( i \) is chosen by individual \( n \), zero otherwise. However, the log of the likelihood is generally used for optimisation, as it has a more tractable form.

\[
\mathcal{L} = \mathcal{L}(\beta; X) = \sum_n \sum_i g_{in} \ln(P_i). \tag{2.3}
\]

The vector of coefficients \( \beta \), hidden in the utility function, is estimated by maximising the log-likelihood function \( \mathcal{L} \). A searching algorithm that optimises the likelihood function performs the actual optimisation, that is, maximises the probability that the observed values are realised (given the parameterisation, the data and that the data may be described by the model). At this point it should be stressed, that in principle, each individual may have his/her own true vector of coefficients, whereas the estimation procedure seeks to assimilate these for ease of estimation \(^3\).

For estimation of coefficients in the discrete choice model, the method of maximum likelihood is applied. An alternative illustration of the method of maximum likelihood (discrete choice between two alternatives) is given in figure 2.1. The figure shows the utility value of two alternatives (\( \triangle \) and \( \square \)) for 7 choice

\(^3\)In technical terms, the individuals are assumed to have Independent and Identical Distributed error terms (IID). This reduces the complexity of the optimisation, as spatial integration is substituted by one-dimensional integrals. For further discussion of this issue see e.g. (Ben-Akiva and Lerman, 1985)
2.2 Model estimation

situations where both alternatives are available. The shape of the point indicate which alternative was chosen. The scale of the vertical and the horizontal axes are the same measuring the utility of the two alternatives; the red dotted line indicate identical values of the two axes \( V(\triangle) = V(\Box) \). The utility value is calculated given the coefficient value and the utility function (in 2.1) – coefficients that optimises the likelihood function, will ensure that \( \triangle \) point are on the upper side of the line, that is \( V(\triangle) > V(\Box) \), \( \Box \) points below where \( V(\triangle) < V(\Box) \). Evidently, the distributions shown in the figure is not an optimum for utility maximisation but is in fact an optimum for random utility maximisation problem\(^4\).

The effect of changes in the \( \beta \) value may be described in two steps. Changes in coefficients for attributes only relating to the \( \triangle \) (\( \Box \)) alternative, will change the location of all points in D/B (A/C) direction. For attributes that relate to both alternatives the relocation are in directions A/C and B/D.

Changing attributes (e.g. improving the service of a train alternative) may or may not alter the outcome, as never 'more than one' of an alternative is chosen.\(^5\) Stated differently, the only information in the data is which alternative was chosen, and not whether it was chosen over the others because it was 'better', 'slightly', 'moderately' or 'extremely much' better than others. Similarly, the scale of utility is ordinal; 2 is better than 1 in the same way as 10,000 is better than 3.

From this graphical version of maximum likelihood estimation, it is evident that the same model specification is estimated irrespective of whether coefficients are scaled by some positive scalar.\(^6\) Due to this, model comparisons are performed between ratios of coefficients and not the coefficients them selves. The most common ratio of coefficients is the value of time given as \( \beta_{\text{time}}/\beta_{\text{cost}} \).

A segmented model where different vectors of coefficients describe sub-samples of the population, is illustrated in figure 2.2. The two segments are identified by colour, the shape of the point indicates the actual choice of alternative. Elements of the blue segment would move down, right (A/B) to form a clearer division on either side of the red line \( (U(\triangle) = U(\Box)) \), indicating change in coefficients. However, elements of the green segment may move up/left (C/D) again indicating change in coefficients, 'opposite' that for the blue segment. It should be noted, that all segments must include a variation of choice of alternative (as

\(^4\)Adding the error term in the utility function corresponds to making the red dotted line 'thicker', wherefore the may be on the correct side of the line.

\(^5\)If the alternative is presently chosen, improving the alternative does not make it 'more chosen'. If not presently chosen, the choice is only altered if the improvement is 'large enough' to make it the best alternative. In terms of economics, (local) non-saturation is not fulfilled, due to the constraint of choice of a minimum of one and no more than one alternative. Choice may not be constructed as 0.3 of alternative 1 and 0.7 of alternative 2, even if this leads to a higher utility value.

\(^6\)This issue is discussed further in section 3.3.
only the difference between utility values are taken into account/measured when model coefficients are estimated).

2.2.2 Properties of the estimator

Describing data by a model involves a set of coefficients ($\theta$) describing the observations, to ‘weight’ the explanatory variables appropriately. An estimator ($\hat{\theta}$), is a random variable (before the sample is drawn from the population), whereas any specific value it may obtain (and we may believe for various reasons) is a number – the latter is generally referred to as the estimate, the former as the estimator. We may speak of the ‘true value’ of the coefficient, as what we aim at in modelling contexts though it is an unobservable and therefore not possible to directly compare with.

Estimators are functions of the data they are estimated from, whether it is the whole population (in principle infinite in size) or, as is generally the case, a sample (subset) of the population. Properties of these estimators are again, subject to the quantity of data as small sample properties holds exactly for any size $N$, whereas asymptotic properties only holds for $N \rightarrow \infty$ (any small sample property hold asymptotically whereas the opposite is not true).

The most obvious concepts to describe an estimator by are distribution, unbiasedness and efficiency. As the estimator is a random variable, we may speak of it as having a distribution. If the distribution is known, we may infer about variation of the estimator (the part that is due to the modelling). An unbiased estimator, on average obtains the ‘true value’, that is $E(\hat{\theta}) = \theta$. The minimum
2.3 Choice context

variance unbiased estimator is termed efficient.\(^7\) Consistency of an estimator (asymptotic property) is the concentration of the distribution at the ‘true value’ as the sample gets larger.\(^8\)

The maximum likelihood estimators are (large sample properties) under suitable regularity conditions:

- consistent
- asymptotically normal
- asymptotically efficient, variance given by the Cramér-Rao bound.\(^9\)

When a model is formulated and estimates found, information on the validity of the model is appreciated. Two types of information can be obtained; whether each of the explanatory variables contribute to explain the variance in the data (coefficient significantly different from zero) and whether some specification is ‘better’ than some other specification in a statistical sense. The first may be tested by the \(t\)-test given by \(\text{mean/ std.error}\), which is asymptotically \(t\) distributed.

Comparisons of different specifications, may be compared e.g. by the likelihood ratio test. This is given by \(-2(L^R - L^U)\), asymptotically distributed \(\chi^2\), where \(L^R(L^U)\) is the likelihood for the restricted (unrestricted) model. Also the \(R^2\) indicates the fraction of variation explained.

### 2.3 Choice context

This section introduces different concepts of choice contexts, whereas a thorough description is postponed to section 3. In the simplest model (the MNL) all individuals are assumed to behave in almost the same way. This is formulated by applying the same coefficients for all individual; possible deviations between individuals are captured by the residual term. The mathematical formulation for this is shown in equation 2.1. The expansion to mixed models corresponds to partitioning the unexplained variation (the \(\nu\)) into a systematic (describable) \(\eta_j\), and an unsystematic part \(\epsilon_j\), while the deterministic part is unaltered \((V = \sum_k \beta_k X_k)\). Formulated as in equation (2.1):

\[
U_{ji} = V_{ji} + \eta_j + \epsilon_j. \tag{2.4}
\]

\(^7\)The minimum variance estimator may be difficult to find as it, at least in principle, needs to be compared to any other estimator. To work around this the Cramér-Rao theorem, that gives the lower bound for the variance, may be employed. The bound is defined by \((\text{var}(\theta) \geq \frac{1}{\mu(\theta^2/\theta^2)})\) where \(\ell\) is the log likelihood, that is the denominator is minus the mean of the Hessian matrix, also referred to as the Information matrix.

\(^8\)For the statistically interested, \(\lim_{N \to \infty} \left(Pr(\theta - q \leq \theta_N \leq \theta + q)\right) = 1\), or in brief notation; \(\text{plim}_{N \to \infty}(\theta_N) = \theta\).

\(^9\)It can be shown that the variance is downward biased \(E(\sigma_N^2) = ((N - 1)/N)\sigma^2\); see (Ben-Akiva and Lerman, 1985).
Assuming shape of distribution of the unsystematic part \((\epsilon_j)\) as e.g. Extreme Value, Type I (Gumbel) corresponds to the Mixed Logit framework\(^\text{10}\), whereas assuming a normal distribution leads to a Mixed Probit model. A special case of the Mixed Logit model is the Random Parameters Logit (RPL), where \(\eta_j\) is formulated as a random variate \(\xi\) multiplied by an explanatory variable, potentially one \(\xi\) for each explanatory variable, while the unsystematic part \((\epsilon)\) is (again) assumed to be Extreme Value, Type I (Gumbel) distributed - though not identical to \(\nu_j\) in (2.1).\(^\text{11}\) Hence, the formulation is as follows

\[
U_{ji} = V_{ji} + \sum_{k' \leq K'} \xi_{k'} X_{jik'} + \epsilon_j, \quad K' \leq K
\]

\[
= \sum_{k \leq K} \beta_k X_{jik} + \sum_{k' \in K'} \xi_{k'} X_{jik'} + \epsilon_j, \quad K' \leq K
\]

\[
= \sum_{k \in K} (\beta_k + \xi_k)X_{jik} + \epsilon_j
\]

(2.5)

The substitution of indices is justified by setting \(\xi_k = 0\) for attributes without an error component added. In matrix notation (2.5) becomes,

\[
U_{ji} = X_{ji}(B + \Xi) + \epsilon_j,
\]

where \(B\) and \(\Xi\) are vectors (diagonal matrices) implying that the error components are independent. Further expansion to (non-diagonal) matrices allows for inclusion of correlations between the elements.

In the RPL formulation the coefficient referred to as \(\beta\) may differ in size from the \(\beta\) referred to in traditional (fixed coefficient model) as they are estimated from different likelihood functions. The \(\beta\) in the RPL is the mean value of a distribution which may not be uni-modal, symmetrical. \(\beta\) in the traditional mode is estimated from a likelihood function based on a number of observations.

The key issue is to uncover the shape of this distribution - whether it is bell-shaped as a normal distribution, uni-modal, skewed or any other (systematic) shape - or, just as importantly, it is not characterised by any well-known (well-defined) shape.

In the literature, an alternative matrix notation is used to describe the model

\(^{10}\)The family of models that may be used to models in principle, any kind of individual behaviour different from the simple 'all individuals behave the same way' (IID) assumed in the MNL model. The 'Mixed Logit Framework' refers to the Mixed Logit Model (McFadden and Train, 2000), Error Components Logit (Ben-Akiva et al., 1993), Logit Kernel (Ben-Akiva et al., 2001), Random Parameters Logit (Ben-Akiva et al., 1993) and Hybrid Choice Model (Ben-Akiva et al., 2002)

\(^{11}\)Assuming a Extreme Value Type I distribution, facilitates estimation by means of the logit framework, often referred to as 'Logit Kernel'. The Extreme Value Type I distribution is described in appendix A.2
2.3 Choice context

where the $\nu$ in 2.1, is specified differently. That is, $\nu_n = F_n \xi_n + \epsilon_n$, where $\xi$ is a $M$-vector of multivariate distributed latent factors and $F_n$ are factor loadings (may be fixed or unknown parameters or a function of attributes). The utility function is as follows,

$$U_n = X\beta + F_n T \xi_n + \nu_n.$$  \hfill (2.7)

This specification is used for estimation purposes and the $\zeta$ is Choleski decomposed into $\xi = T \zeta$, where $\zeta$ is a matrix of independent factors, and $TT'$ is the covariance matrix of $\xi$. Hereby the covariance between the utility of alternatives, is given by

$$\text{Cov}(U_n) = F_n TT' F_n' + (g/\mu^2) I_n.$$  \hfill (2.8)

The 'Mother Logit Theorem' states that any choice model can be approximated by a model that takes the form of a (standard) logit. The result is reproduced in Revelt and Train (1998) but originally due to McFadden (in 1975).

Recent research (McFadden and Train, 2000), has shown that models of complex nesting structures and models of complex stochastic structures are similar as any tree structure may be approximated by a ML specification. Yet it is realised, that the number of different nesting structures (no restrictions on alternatives) rises extremely fast with number of alternatives involved. Allowing for alternatives to appear in more than one nest (or to be chosen via different paths through the tree), only worsen this problem. As for the mixed models the issue of constructing the models are either diverted to obtain a good model fit with a few additional terms or to gain understanding of what drives the individual’s choice process (tastes) which may require a (larger) number of terms; these possibly correlated.

2.3.1 Extending complexity of models

The simple and very tractable formulation of the logit model has been used as the base for many models. What may be termed as the simplest extension, is that of data segmentation. Setting up different models (same model type, different utility functions) for fraction of the data resembles assuming that choices are based on different reasonings or preferences for (groups of) individuals. This a priori segmentation is typically tested just as a comparison to the unsegmented model and not compared to other alternative segmentations of data. Data segmentation is a special case of latent classes (class membership is deterministic) and of the mixed logit model with a discrete distribution (degenerate for each data segment).

Mixed logit models can be seen a the general case of a segmented model, where each individual has a specific model formulation (though the included variables are the same for all individuals) with the individuals own relative valuation of each of the attributes of the alternatives. Such a model would account for inter person variations, i.e. that different individuals may choose differently given the same choice context. On the other hand, intra person variations, that the same
CHAPTER 2. THE BEHAVIOUR OF CHOICE

Alternatives | Identified (Order) | Heterogeneous # Unknowns | Corr. Heterogeneous # Unknowns
---|---|---|---
2 | 0 | 3 | 4
3 | 2 | 4 | 7
4 | 5 | 5 | 11
5 | 9 | 6 | 16
6 | 14 | 7 | 22
7 | 20 | 8 | 29
8 | 27 | 9 | 37
9 | 35 | 10 | 48
10 | 46 | 11 | 58

Table 2.1: Number of identified coefficients for each case

individual chooses differently in the same choice context, but at different times is yet not accounted for (the latter also referred to as between-time variation).

A mixed logit model with a continuous distribution over all data records (i.e. over individuals and their repetitive choice situations) accounts for both inter and intra person variations.

2.3.2 Identification of models

Before estimating the coefficients of a given model specification identification of the model should be ensured - as coefficient estimates may be obtained even for models that are not identified. Identifiability of a model is a matter of whether the proposed model specification can be estimated, that is, whether estimators exists for each and all of the included model coefficients.

Model identification can be ensured by the order (maximum number of coefficients that may be estimated) and rank (the actual number of coefficients that may be estimated) conditions (Bunch, 1991). The order condition is a necessary condition for identifiability whereas the rank condition is a sufficient condition for identification. Further, identification of Logit Kernel models is described in Walker (2000). If a normalisation must be included in the model to ensure identification of the model (in addition to that of the Gumbel variance set to 1), the positive definiteness condition may be introduced to check the validity of the normalisation12. The number of identified parameters can be seen from table 2.1. Generally, all coefficients in a heterogeneous model13 are not identified. It is stressed that these are maximal number of identified coefficients – cases may occur where the actual number of identified coefficients is lower.

12Walker (2000) shows that a previously applied normalisation was not valid due to this criterion.
13See section 3.7 for a description of these models.
2.3 Choice context

In a fixed coefficients logit model the number of coefficients to be estimated is \( J + 1 \), where \( J \) is number of coefficients \( \beta \), in addition to the scale \( \mu \). In this case \( s = J(J - 1)/2 - 1 \) alternative-specific coefficients are identified due to the order condition (see Walker (2000)), hence identification is not a problem. Adding random coefficients (normal distribution) the number of coefficients to be estimated increases to \( 2J + 1 \), if correlation between the coefficients is not allowed, or to \( 2J + 1 + J(J - 1)/2 = J(J/2 + 3/2) + 1 \) if a full correlation structure between the coefficients \((\beta)\) is invoked\(^{14}\).

Future research will, I believe, improve the understanding of the possibilities and limitations of the mixed models and models of non-independent alternatives. As understanding of the models have remarkably improved this has exposed new gaps; ‘comprehendible’ models (cross-nested or simpler tree structure) versus mixed models, aim at model fit versus understanding of individuals choice formation, how many distributed terms to employ in an model, which distribution(s) to apply, should the distributions be correlated?

The following section gives a literature review of (most of) existing model variants applied for mode choice, by now. From the classical MNL over rising complexity tree structures to latent variable models and mixed model formulations.

\(^{14}\)In order to estimate the parameters of the distribution of the coefficients (by Maximum Simulated Likelihood, described in section 4.6.1) an even higher number of parameters may be involved, depending on the distribution assumed for the coefficients.
Chapter 3

Existing models

3.1 Background

Daniel Bernoulli was the first (1738) to announce the concept of decision theory, to explain people’s non-linear value of money. The theory was further revived and developed in the 1920’s by Frank Ramsey and subsequently by Neumann and Morgenstern (1944). Savage (1954) formalised the set of decision theory axioms, providing much debate and controversy, but setting the benchmarks.

Decision Theory defines the concept of rationality with reference to decision making, and proves that, assuming acceptance of the four main axioms, as a decision maker one should always choose the alternative which maximise the expected (and perceived) utility. The main axioms are:\footnote{As the case is discrete choice analysis, local non-saturation is left out.}

- Ordering - alternatives can be put in (unambiguous) order of preference.
- Transitivity - if A is preferred to B and B to C, then A is preferred to C.
- Dominance - if A is as good as B in every respect and A is better than B in at least one respect, then one should prefer A to B.

A discrete choice is the choice of exactly one alternative from a finite set of alternatives. To operationalise this, behavioural theory is added including utility to measure the 'worth' of alternatives. The theory of marginal utility was developed independently William Stanley Jevons (England 1871), Karl Menger (Austria 1871) and Leon Walras (Switzerland, 1874) and formed the foundation for all subsequent economic science. At the early stage utility was thought to be \textit{cardinal}, i.e. that a utility of 6 was three times better that a utility of 2, whereas later the concept of \textit{ordinal} utility has been successful.\footnote{On an ordinal scale of utility, 3 is better than 2, in the same way as 10,000 is better than 3.}
This extends the choice situation to be choice of the best alternative measured on an arbitrary scale of utility. In general, the choice of alternative $i$ from a set of alternatives $C$, where the utility of the alternative is subdivided into a measurable part $V$ and an error term $\epsilon$, can be described by the following

$$P(i) = P(V_i + \epsilon_i \geq V_j + \epsilon_j, j \neq i)$$

(3.1)

$$= P(V_i - V_j \geq \epsilon_j - \epsilon_i, j \neq i).$$

The set of alternatives $C$ constitutes of (for mode choice) all possible modes $A_i(i = 1, \ldots k)$ that an individual can apply for a given trip, that is,

$$\bar{C} = \{A_1, \ldots, A_k\} \text{ where } P(A_i) \geq 0.$$  

(3.2)

Though, it can be argued that this set should be reduced to relevant alternatives. Relevant alternatives are those considered by the individuals and therefore described by a positive probability for being chosen (if an alternative is not considered by the individual it will never be chosen by the individual). This set is given by

$$C = \{A_1, \ldots, A_n\} \text{ where } P(A_i) > 0,$$

(3.3)

where $n < k$ and $C \subseteq \bar{C}$. The reduction has no impact on the estimators but merely reduces the preparatory work (describing alternatives). Different assumptions of distribution of the error term and the specification of the utility function lead to the well-known models described in the following chapter. Attention is directed towards models applied for mode choice, though applications of discrete choice models are found within transport assignment, market research, psychology, biometrics, agriculture etc. The list of models presented is not guaranteed to be exhaustive of models documented in the literature, though quite a number of models are described.

The following is ordered after ‘rising model complexity’. That is, starting with models with a maximum of one stochastic term (the residual term) followed by the general description of (most) known choice models, the GEV formulation. Subsequently, models allowing for similarities between alternatives (less trivial tree structures) – an important propagation and alternative to the IIA restriction, are described. Challenges of the single stochastic term’s distribution (the IID restriction) add heteroscedasticity to the model complexity. The section on mixed models concludes the description.

### 3.2 Models

The foundation of measuring the impact of any change in attributes of alternatives is the theory of utility. The individuals perceived worth of an alternative is quantified by some number of attributes each multiplied by a coefficient and a residual term to capture deviations from the average utility. Specifically, for individual $j$

$$U_{ji} = Bh(X) + \epsilon_j,$$

(3.4)
3.3 The Logit model

where \( h(\cdot) \) is some function of the vector of attributes \( X \) multiplied by the vector of coefficients \( B = \{\beta_j\} \). Or, assuming that the utility is described by a linear function of the attributes,

\[
U_{ji} = \sum \beta_j X_{ji} + \epsilon_j
\]  

(3.5)

is used to describe the (indirect) utility for alternative \( i \).

3.3 The Logit model

The most generally recognised model and the reference model for most (if not all) recent model developments, within the area of stochastic utility-based models for description of discrete choice, is the logit model. The model is referred to as the logit model, the multinomial logit or simply the MNL model, interchangeably.

Estimation of which alternative is preferred over others, involves exhaustive evaluation of in principle all alternatives’ utility. Generally, alternatives utility cannot be completely described, setting out the need for adding a (distributed) residual term. This residual is not in itself interesting, merely the (distribution of the) maximum of the random utilities (which includes the residuals). Given some not particularly restrictive, assumptions on the residuals, the shape of the maximum over the alternatives is known.

The most common of models, the (multinomial) logit model (MNL), can be derived in this way, based on additional of some restrictive assumptions of independence between alternatives (IIA) and independence between individuals (IID). The attractiveness of the model is easily seen from the formulation below,

\[
P(i) = \frac{\exp(V_i)}{\sum_j \exp(V_j)},
\]  

(3.6)

where \( i, j \) are alternatives. A graphical representation of the model is a tree structure where alternatives are put directly under the root. In the case of just two alternatives, the formulation reduces to the odds-ratio. Estimation of

\(^3\)Assuming a exponential, normal, Weibull or Gumbel distribution for the residuals in the utility for some alternative, leads to the Extreme Value, Type I distribution, also known as the Gumbel distribution. On the other hand, a Uniform distribution leads to the less known Extreme Value, Type III, see e.g. (Kotz and Johnson, 1982-99) for properties of maximum of stochastic variables.

\(^4\)The IIA property (Independence of Irrelevant Alternatives) states that the rate of substitution between two alternatives is unchanged by the addition of an extra alternative/ removal of an alternative. It is a consequence of Luce’ choice axiom of 1959, stating that \( P(i|\tilde{C}_n \subseteq C) = P(i|C), i \in \tilde{C}_n \subseteq C \), (Luce, 1959).

\(^5\)Independent and Identically Distributed error terms is a technical assumption to ease the derivation of the models from stochastic utility (the integral over \( n \)-dimensional space is substituted by \( n \) one-dimensional integrals).
model coefficients are based on maximum likelihood estimation, that is in practice maximisation of the log likelihood function \( \ell \).

At this point it should be noted, that coefficient estimates of \( \beta \) are scaled by a logit scale factor \( \mu \) related to the variance of the error term \( \epsilon \) by \( \mu^2 = \pi^2/6\epsilon^2 \). When comparing models this must be borne in mind.\(^6\)

Various types of information on the alternatives can be incorporated into the model. Examples are rank ordering of alternatives (e.g. choice of number of cars in a household), which can be modelled by use of the ordered logit model. Similarity between groups of alternatives may be incorporated as well, where the model formulation depends on how/whether these similarities can be constructed by elements related to non-overlapping (sub) groups or not. These models are described in later sections.

**The Incremental Logit Model**

The incremental version of the logit model can by used for assessing the effect of changing the service levels of one/some variables without specifying all levels, but merely changes in levels.

\[
p'_k = \frac{p^0_k \exp(V_k - V^0_k)}{\sum_k p^0_k \exp(V_k - V^0_k)}, \tag{3.7}
\]

where \( p'_k \) is the new proportion of trips using mode \( k \). The model has later been referred to as pivot point model.

Models applied for mode choice today are by far more complex than those just described. During the 1970’es a theoretical framework was set up describing a very broad class of models. This is the topic of the following section.

### 3.4 The GEV family

The description of models for discrete choice were systematised by McFadden in 1974/78 (see the well known paper McFadden (1978)). The term *Generalised Extreme Value* (GEV) evolved, and members of this family were described as well as their properties. In the following sections known members of the family are summarised and references are provided (most of the known models are members of the GEV family).

\(^6\)For logit models a logit scale \( \mu \) is introduced (all coefficients are multiplied by it) as the choice of alternative is inferior to the scale of the utility. Mathematically, \( P(V_i + \epsilon_i \geq V_j + \epsilon_j) = P(\mu V_i + \mu \epsilon_i \geq \mu V_j + \mu \epsilon_j) = P(\mu(V_i - V_j) \geq \mu(\epsilon_j - \epsilon_i)) \). The coefficients in the utility function (\( \beta \)) are not separable from the scale \( \mu \). For the sake of identifiability, the variance of the residual term is set to one, whereby \( \mu \) gives the variance of the residual term.
3.5 Models with correlated alternatives

For members of the GEV, choice probabilities can be described by

\[ P(i) = \frac{\exp(V_i + \ln G_i(y_1, \ldots, y_N))}{\sum_j \exp(V_j + \ln G_j(y_1, \ldots, y_N))} \]  

(3.8)

where the probability for \( P(i) \) for alternative \( i \) described by (systematic) utility \( V_i \). \( G_i \) is the partial derivative of \( G \) with respect to the \( i \)th of the descriptive variables \( y_1, \ldots, y_N \).

The family of GEV Models cover all model formulations that satisfies the following properties.

- \( G \) is non-negative.
- \( G \) is homogeneous of degree \( \mu \geq 0 \)
- \( \lim_{y_i \to \infty} G = \infty \), \( \forall y_i \).
- The \( l \)th partial derivative is non-negative (non-positive) if \( l \) is uneven (even)
- \( y_1, \ldots, y_N \geq 0 \).

The last criteria is a technicality, simply satisfied by letting \( y_i = \exp V_i \). In a later version (Ben-Akiva and Lerman, 1985) this is not included, though it seems to be necessary in order to assure symmetrical second derivatives (Daly, 2001b). By application of Euler’s theorem\(^8\) the GEV form becomes

\[ P(i) = \frac{e^{V_i}G_i}{\mu G}, \]  

(3.9)

from which known models are obtained by substitution of the relevant function \( G \). An important implication of the GEV specification is, that if a model can be specified as a GEV then it is consistent with Random Utility Maximisation (RUM) theory.

3.5 Models with correlated alternatives

The IIA condition that has been argued to be unrealistic by the red bus-blue bus problem, is relaxed in the following models, implying a level of correlation or dependence between alternatives. First, relaxations through more complex tree structures (multilevel or even cross-nesting) are presented. Then follows models that takes advantage of ranking and then models of more mistily defined preferences of alternatives. This dependence or correlation, between alternatives, may

---

7In the original formulation homogeneity of degree 1 was required, though an unpublished paper by Ben-Akiva and Francois, apparently states that homogeneity of any positive degree is sufficiently. The result is subsequently published in (Ben-Akiva and Lerman, 1985).

8Euler’s theorem states \( \mu G = \sum_{j \in J} y_j G_j(y_1, \ldots, y_N) \)
CHAPTER 3. EXISTING MODELS

in some cases (e.g. the PLC, dogit) easiest be seen through the cross elasticities, which states the rate of change in the probability of choosing one alternative given the rate of change of some attribute of another alternative, see section 3.8.

3.5.1 The Nested Logit Model

Probably, the most well known model that may handle non-independent alternatives, is the Nested Logit model (or Hierarchical Logit Model), or NL model for short. The Nested Logit model applies to situations with at least three alternatives, where some (but not all) of these share a common unobserved characteristic, e.g. type of mode, level of comfort. Furthermore, the NL and the MNL models are the most widely used models in practical applications.

In the Nested Logit Model the set of alternatives is divided (and sub-divided) into exclusive groups (nests) where some aspect only pertains to members of that particular group. Graphically, this is represented as nests of alternatives under the root and may in principle have any number of levels, though limited by the number of alternatives. The construct is in opposition with the IIA. Estimation is performed by estimation of lower nests and upwards in the tree by handling nests as 'combined' alternative, which’ value is measured by the log sum. The NL model was developed in parallel by Daly and Zachary (1978) independently of the work by Williams (1977). For estimation purposes there exist two different formulations; the non-normalised version (NNNL) proposed by Daly and the normalised (utility maximising, UMNL) version by McFadden 9.

A motivation for the model construct is given in (Ben-Akiva and Lerman, 1985). If the alternatives have a common observable characteristics as well as unobserved characteristics the general utility function may be written as follows

\[ U_{dm} = \tilde{V}_d + \tilde{V}_m + \tilde{V}_{dm} + \tilde{\epsilon}_m + \tilde{\epsilon}_d + \tilde{\epsilon}_{dm}, \] (3.10)

By introducing the two extra error terms (\(\tilde{\epsilon}_d, \tilde{\epsilon}_m\)), it can easily be seen that alternatives has become correlated (dependent), as \(\text{cov}(U_{dm}, U_{dm'}) = \tilde{\epsilon}_m\) respective \(\text{cov}(U_{dm}, U'_{dm}) = \tilde{\epsilon}_m\).

If either \(\tilde{\epsilon}_d\) or \(\tilde{\epsilon}_m\) can be neglected, a tree structure can be formed, with the lower level corresponding to the neglected error term (choice of \(d\) at the lower level if \(\tilde{\epsilon}_d\) is neglected). In practice, as the variance is not zero, the least correlated level of the choice is at the lowest level in the tree. This is verified in the estimation of the model, by means of the scaling parameter where \(\mu_d \leq \mu_m\).

The joint logit is a special case of the Nested logit model, where \(\mu_d = \mu_m\).

In Daly (1987) an estimation procedure for a nested logit model is proposed.

9A recent paper by Orutzar (2001) gives an overview of the practical and theoretical publications related to the development of the NL.
3.5 Models with correlated alternatives

The Non-Normalised Nested Logit (NNNL) model given by

\[ P(n) = P(n|m)P(m), \quad P(n|m) = \frac{e^{V_n}}{\sum_{n' \in N_m} e^{V_{n'}}}, \quad (3.11) \]

\[ P(m) = \frac{e^{\mu_m \Gamma_m}}{\sum_{m'=1}^{M} e^{\mu_{m'} / \Gamma_{m'}}}, \quad \Gamma_m = \ln \sum_{n' \in N_m} e^{V_{n'}/\mu_m}. \]

NNNL is not consistent with utility maximisation as noted in Daly (1987); Koppelman (1998). In McFadden (1978) the Utility Maximising Nested Logit (UMNL) model given by.

\[ P(n) = P(n|m)P(m), \quad P(n|m) = \frac{e^{V_n/\mu_m}}{\sum_{n' \in N_m} e^{V_{n'/\mu_m}}}, \quad (3.12) \]

\[ P(m) = \frac{e^{\mu_m \Gamma_m}}{\sum_{m'=1}^{M} e^{\mu_{m'}/\Gamma_{m'}}}, \quad \Gamma_m = \ln \sum_{n' \in N_m} e^{V_{n'/\mu_m}}. \]

The difference between the two models is the normalisation by the logsum in UMNL. Recently, the differences between the two estimation approaches has been narrowed as estimates from one approach may be transformed to estimates of the other approach (Koppelman and Wen, 1998; Daly, 2001a).

**The Extended Nested Logit model**

The term Extended Nested Logit (ENL) emerges in Vovsha (1996) where the NL is reformulated for use in explaining the CNL model (described in section 3.5.3). Basically, the model is a NL with an extra index added, which explicitly states through which nest \( m \) in the tree the alternative \( i \) is chosen, though each alternative is only present under one nest. The formula is

\[ P^E(i, m) = \frac{e^{V_{im} + (\mu - 1) \ln \sum_{k \in C_m(i)} e^{V_{km}}}}{\sum_{j \in \mathcal{C}} e^{V_{jm} + (\mu - 1) \ln \sum_{k \in C_m(j)} e^{V_{km}}}}, \quad (3.13) \]

where \( V_{im} = \beta_{im}^0 + V_i \) that is, an alternative and characteristic (nest) specific constant is added to the utility for alternative \( i \). \( \mu \) is the logsum parameter.

**The Single Element Nested Logit Model**

A variant of the nested logit is the Single Element Nested Logit (SENL). The SENL is a tree structure where nests directly under the root correspond to alternatives available for each data source. The subsequent levels forms a nest structure similar to a Nested Logit for each data source. A nesting parameter
is added for each nest under the root (data source) – one is constrained for the
sake of identification. To account for different levels of variance between data
sources at lower levels, nesting parameters are added as in any other NL model.

The SENL approach is generally applied to increase to amount of data a model
is based upon, by reusing previously collected data, due to the high cost of
data collection. Furthermore, the approach applies for models including non-
deterministic behaviour\footnote{These model variants are described later in section 3.6.}. The model is further described in (Bradley and Daly, 1991).

3.5.2 The Joint Logit model

The joint logit model allows alternatives to have common observed characteris-
tics, which can partly be described by a systematic (observed) part of the utility
and partly by an unobserved part of the utility (the error term). The model is
constructed as a hierarchical model; for simplicity a two level model \((d, m)\) is
described here. The utility function for the Joint Logit Model is as follow

\[
U_{dm} = \tilde{V}_d + \tilde{V}_m + \tilde{V}_{dm} + \epsilon_{dm},
\]

where \(\tilde{V}_d (\tilde{V}_m)\) is the systematic utility related solely to the characteristic \(d (m)\),
whereas \(\tilde{V}_{dm}\) is the (systematic) utility connected to the synergy between these
characteristics, while \(\epsilon_{dm}\) is the residual related to both characteristics \(d, m\).

Marginal probabilities are described similarly to the MNL model. The model
is further described in Ben-Akiva and Lerman (1985). Had the commonality of
the alternatives been unobserved, then the describing model would have been
the Nested Logit Model described in the following section.

3.5.3 The Generalised Nested Logit

The Generalised Nested Logit (GNL) model is a generalisation of the CNL and
PCL model, allowing for the liberal cross-nesting structure in the CNL model
and the flexibility of the PCL model. The model is formulated in Wen and
Koppelman (2001), with comparisons to the CNL, PCL, OGEV and DP mod-
els and further (GenMNL, FNL, CCL, FNL, Dogit, PLC, HMNL, COVNL) in
Koppelman and Sethi (2000).

Alternatives are arranged in a nesting structure where each alternative may
enter more than one nest (weighted by an allocation parameter), as long as the
'total number' of each alternative is 1. The probability of choosing an alter-
native is the product of probability of choice of the alternative given a nest,
and summed over all nests that include the alternative. The formulation is as
3.5 Models with correlated alternatives

follows;

\[ P(i) = \sum_m P_{i|m} \cdot P_m \]

\[ = \sum_m \left[ \frac{\alpha_{im} e^{V_i}}{\sum_{j \in N_m} (\alpha_{jm} e^{V_j})^{\mu_m}} \cdot \frac{\left( \sum_{j \in N_m} (\alpha_{jm} e^{V_j})^{\mu_m} \right)^{\mu_{m'}}}{\sum_{m'} \left( \sum_{j \in N_{m'}} (\alpha_{jm'} e^{V_j})^{\mu_{m'}} \right)^{\mu_{m'}}} \right], \]

where \( N_m \) is number of alternatives in nest \( m \), \( \alpha_{im} \geq 0 \) is the proportion of alternative \( i \) assigned to nest \( m \) adding to 1 over nests (each alternative accounted for exactly one time), while \( 0 < \mu_m \leq 1 \) is a logsum (dissimilarity) parameter. The constraint \( \sum_m \alpha_{im} = 1 \) ensures that each alternative occur a 'total' of one time.

The Cross-Nested Logit Model

The Cross-Nested Logit (CNL) model appears in two version by Vovsha - with different solution algorithms (Vovsha, 1996, 1997). The model is formulated as

\[ P(i) = \sum_m \frac{\alpha_{im} e^{V_i}}{\sum_{k \in C} \alpha_{km} e^{V_k}} \sum_{m'} \left( \sum_{k \in C} \alpha_{km} e^{V_k} \right)^{\mu_{m'}}, \]

where the weights of alternatives in each nest must satisfy \( \sum_k \alpha_{km}^{\mu_{m}} = 1 \) for nest \( m, k \) alternative. The number of levels in the nesting structure is confined to two.

The CNL model is a special case the GNL which can be seen by letting the structural parameters (logsums) be equal \( \theta_m = \theta \). Variants of the CNL exists, the Link-Nested Logit (LNL) by Vovsha and Bekhor (1998), Small’s OGEV model (see section 3.5.4), a version by Papola defined by \( G(y_1, \ldots, y_J) = \sum_m \left( \sum_{j \in C} \alpha_{jm} e^{V_j} \right)^{\mu_m} \) and a version by Ben-Akiva and Bierlaire with \( G(y_1, \ldots, y_J) = \sum_m (\sum_{j \in C} \alpha_{jm} y_j^{\mu_m})^{\mu_{m'}}. \)

The Paired Combinatorial Logit model

The Paired Combinatorial Logit (PCL) model was introduced in Chu (1989)\(^{11}\) and reappeared in Koppelman (2000)\(^{12}\). Applications of the model has been hard to find, probably due to the lack of software that is capable of handling the model without extensive programming. The model in principle allows correlation between any pair of alternatives. The 'joint aspects' of alternatives, are added as similarity coefficients \( \sigma_{ij} \in (0, 1) \) where the value 1 is virtually identical alternatives, while 0 corresponds to the IIA property. The choice probabilities

\(^{11}\)Apparently, the model was described in Chu’s unpublished Ph.D. thesis (1981).

\(^{12}\)The paper was previously presented at the TRB, 1996.
are described by

\[
P(i) = \frac{\sum \exp \left( \frac{V_i}{1-\sigma_{ij}} \right) \left( \exp \left( \frac{V_i}{1-\sigma_{ij}} \right) + \exp \left( \frac{V_j}{1-\sigma_{ij}} \right) \right)^{-\sigma_{ij}}}{\sum_{k=1}^{n-1} \sum_{m = k+1}^{n} (1-\sigma_{km}) \left( \exp \left( \frac{V_k}{1-\sigma_{km}} \right) + \exp \left( \frac{V_m}{1-\sigma_{km}} \right) \right)^{1-\sigma_{km}}}
\]

\[= \sum_{j \neq i} P(ij)P(i \mid (ij)). \tag{3.17} \]

The PCL model can be obtained from the GNL by formulation of a nesting structure where each nest consists of two alternatives and constraining the allocation parameters to be equal \(\alpha_{ij} = \alpha\). This too eases the task of estimating the model coefficients.

### The Product Differentiation Model

The Product Differentiation (PD) Model was proposed within the framework of a market with a number of (to some extent) similar products. Similarities are handled in terms of dimensions that characterise attributes of the products. Transferred to terms of transport modelling, dimensions could include mode and location choice, number of cars etc. The model is described in Bresnahan et al. (1997).

The PD model allocates each alternative to a nest along each of a set of pre-selected dimensions, logsum parameters constrained to be equal for each nest for each dimension. Formulated below;

\[
P_i = \sum_{d \in D} \left[ \alpha_{id} \frac{e^{V_i/\mu_d}}{\sum_{k \in d} e^{V_k/\mu_d} \sum_{d' \in D} \left( \sum_{k' \in d'} e^{V_{k'}/\mu_{d'}} \right)^{\mu_{d'}}} \right], \tag{3.18} \]

where \(\mu_d\) is logsum and \(\alpha_{id}\) is proportion of the alternative allocated to dimension \(d\). The PD is a special case of GNL where all alternatives have the same allocation to each dimension and nests along each dimension have the same logsum parameter (Wen and Koppelman, 2001).

### 3.5.4 The Ordered Generalised Extreme Value model

The Ordered Generalised Extreme Value (OGEV) model described in Small (1987). Different cross-elasticity between alternatives in the same nest is allowed, hereby allowing for different cross-elasticities over pairs of alternatives. Only pairs of alternatives adjacent in order are allowed membership of the same nest. As alternatives are allowed membership of more than one nest (weighed, such the number of alternatives adds to one) with different logsum parameters
3.5 Models with correlated alternatives

Attached, different cross-elasticities are possible. The model is formulated as

\[ P_i = \sum_{m=1}^{i+L} P_{i|m} P_m \]

\[ = \sum_{m=1}^{i+L} \left[ \frac{(w_{m-i}e^{V_i})^{1/\mu_m}}{\sum_{j \in N_m}(w_{m-j}e^{V_j})^{1/\mu_m}} \sum_{s=1}^{j+L} \left( \frac{\sum_{j \in N_m}(w_{m-j}e^{V_j})^{1/\mu_m}}{\mu_m} \right) \right], \]

where \( w \) is allocation weights for alternatives in the nests each containing a maximum of \( L \) (contiguous) alternatives.

3.5.5 Ordered models

If an unambiguous order of alternatives can be formed, an ordered model may be applied. As the search is for a local optimum instead of a global optimum, estimation time is reduced. Example of application is modelling of car ownership, i.e. whether a family will own one car, two cars, etc. which are clearly ranked. The ordered logistic model is described in Ben-Akiva and Lerman (1985). Multiplying the probabilities of moving to the next state in the order, forms choice probabilities, i.e.

\[ P(i) = \left( 1 - \frac{1}{1 + e^{-(V_i - V_j)}} \right) \prod_{j=1}^{i-1} \frac{1}{1 + e^{-(V_i - V_j)}}, \]  

3.5.6 The Covariance Heterogeneity Nested Logit model

The MNL assumes identical variance of the residual term, the NL relaxes this by letting alternatives in different nest differ in variance. In the COVariance heterogeneity Nested Logit (COVNL) model proposed by (Bhat, 1995) alternatives within the same nest are allowed heterogeneity. The utility function is reformulated with the residual term divided into an alternative-specific part and a nest-specific part. For a two-level tree structure with three alternatives (one nest containing \( b \) and \( c \)), utility is formulated as

\[ U_a = V_a + \epsilon_a, U_j = V_j + \epsilon_{next,j}, j \in \{b, c\}. \]  

The probabilities of the alternatives are given by,

\[ P_a = \frac{\exp(V_a)}{\exp(V_a) + \exp(\theta \ln \Gamma)} \Gamma = \frac{\exp(V_a)}{\exp(V_a) + \exp(\theta \ln \Gamma)} \exp(V_a/\theta) + \exp(V_c/\theta) \]

\[ P_j = \frac{\exp(V_j/\theta)}{\exp(\theta \ln \Gamma)} \exp(V_j)/\theta, \]

If the \( \theta \) (0 < \( \theta \) ≤ 1) are constrained equal over individuals, the NL model is obtained. The model is on closed form (directly estimable).
CHAPTER 3. EXISTING MODELS

The amplification of the COVNL compared to the NL is somewhat limited to nests of two alternatives. In the NL, a nests containing three alternatives \((a, b, c)\) may be further split into a nest of two alternatives \((a, b)\) and one separate alternative \((c)\), whereby the different covariance between \(a, c\) and \(b, c\) are ensured by the scaling. Similarly, for larger nests. The model is clearly a special case of the Mixed Logit model, described section 3.6.2.

3.5.7 The Recursive Nested Extreme Value model

Hierarchical models are confined by the involved tree structure. Extra alternatives can be added to nests in a matter that ensures cross-correlation to any (one, or nest of) alternative, but not to all other alternatives as the model is limited by it’s own tree structure. Recently, a generalisation of among others, the NL model has been proposed, called the Recursive Nested Extreme Value (RNEV) model, see Daly (2001b). The model is formulated around a tree structure where each (elementary) alternative may be chosen based on different paths (sequences of nests) through the tree, or stated differently, different and complex (more than two-level) competitive issues may be modelled. The probability of choosing an alternative is the sum over different paths through the tree.

\[
p_e = \sum_{s \in S_e} \prod_{k \in S, d \in s(k)} \lambda_{kd} Y_k^{\mu_d/\mu_k} / Y_d
\]  

(3.23)

where \(S_e\) is the set of sequences of nodes to alternative \(e\), \(s(k)\) nodes under nest \(k\) and the \(\mu\)’s are parameters. \(\lambda\) is explained as 'alternative-nest-specific-constants'. They corresponds to the \(\lambda\) in the PCL model, without the PCL-restriction of adding to one. In the RNEV model a constraint of the \(\lambda\)’s is required to avoid over-specification. The parameters must satisfy three constraints. First, \(\lambda, \mu > 0\); second, \(\mu\) non-decreasing for lower levels (along a sequence) in the tree and finally, \(\mu\) is constant over (elementary) alternatives.

Substituting \(V_k(s) = \log(\lambda_{kd} Y_k^{\mu_d/\mu_k})\), \(d = s(k)\) into (3.23) gives

\[
p_e = \sum_{s \in S_e} \left[ \exp \sum_{k \in s} (V_k(s) - \log \sum_{s(h) = s(k)} \exp V_h(s)) \right]
\]  

(3.24)

This formulation shows the close relation to the NL model. Moreover, the specification of the summation indicates a rigorous tree structure where, in principle, any number of levels in the tree is allowed. The RNEV is shown to generalise the OGEV, PD, GNL, TL models and hence, all special cases of these. Moreover the RNEV is a GEV model, whereby properties of the GEV apply to the RNEV and it’s special cases.

Estimation of the RNEV model, can be performed by modification of the algorithm for NL estimation, though applications remains to be seen in the literature. Similarly, the model variants that RNEV generalises, may be estimated this way – hereby, enabling the use of these variants.
3.5 Models with correlated alternatives

The Network GEV model

The issue of describing correlations between alternatives by means of complex tree structures has recently been addressed by Bierlaire (2002). The formulation is termed Network GEV, which is shown to be a GEV model, hence all properties of GEV are inherited. The idea of the model is that cross-nesting is allowed throughout the multi-level tree, as long as circular loops are avoided. Alternatives (at the lowest level of the tree) may be chosen through different paths of the tree (fraction of each alternative measured by weights $\alpha$) and scales attached to each node, rising downwards through the tree. The Network GEV is a counterpart to RNEV (Daly, 2002).

3.5.8 The C-Logit model

This model was implemented for a route choice problem with some routes overlapping, see Cascetta et al. (1996). The model accounts for the overlap of paths – or correlation between alternatives, by measuring the extent of the overlap and reducing the utility of the affected alternatives by this amount. The overlap is termed the 'Commonality Factor' (CF). In technical terms, the deterministic part of the utility of an alternative (route) is altered to $V_k = V_k - CF_k$, where $V_k$ is (traditional) utility of the route and $CF_k$ quantifies the impact of the overlap (path $k$). This is incorporated into the MNL probability to form the C-Logit model. Cascetta et al. (1996); Ramming (2002) list four different formulations for how to calculate the 'commonality factor' (CF). These are all proportional to the (log of) the length of the overlap.\textsuperscript{13} The author is not aware of applications on mode choice models.

3.5.9 The Probit model

As opposed to most of the above mentioned models with Extreme Value, Type I (Gumbel) distributed residual terms, the probit model has normal distributed residual terms. Correlation between any pair of residual terms, hence alternatives, can be modelled hereby, any common aspects for some alternatives can be modelled along with the characteristics specific to a single alternative. Ultimately, a full covariance structure can be employed, which adds to the estimation complexity.

\begin{equation}
P(i) = \int_{V_1 = -\infty}^{V_i - \epsilon_i} \ldots \int_{V_N = -\infty}^{V_N + \epsilon_i} \Phi_{\epsilon_1, \ldots, \epsilon_N} (\epsilon_1, \ldots, \epsilon_N) de_N \ldots de_1
\end{equation}

where $\Phi_{\epsilon_1, \ldots, \epsilon_N}$ is the simultaneous distribution function of the $N$-dimensional normal distribution. Estimation of the model requires the use of simulation as the model (3.25) is not on closed form.

\textsuperscript{13}The close variant, the Path Size Logit, also attempts to measure the route overlap/correlation between alternatives, though applied to route choice models, see Ramming (2002)
Until the early 1980-s estimation of the probit model was limited to three dimension (binary probit, 3-dimensional model hidden in binary models by considering differences between alternatives\footnote{The Technique is known as Clark’s approximation.} rather than the alternatives themselves). As the field of quadrature and numerical simulation (and the computer power) has grown, high dimension probit models has become within reach of a modeller. Estimations techniques are treated in section 4.

The flexibility of the covariance structure leads to that any change due to addition/ removal of an alternative easily can be modelled, which suggest frequent use of the model. Nonetheless, the use is fairly limited when mode choice is the subject. Considering assignment, the probit model is one of the more frequently used models, in particular when there is extensive route overlapping (correlation between alternatives).

The Tobit model

A modified version of the probit model exists where negative values of the dependent variable are censored, the error term inverse normal with a zero mean. The model is known as the Tobit model (or Tobin’s probit). The model is formulated below.

\[ y_i^* = \beta x_i + u_i \text{ where } y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}, u_i \sim IN(0, \sigma^2) \]  

Applications of the model in relation to traffic modelling are given in (Amemiya, 1984).

3.5.10 Latent Preference models

The class ‘Latent Preference models’ is emerging (Ben-Akiva, 2002). The class includes models where the probability for choosing an alternative is in part described by the relative utilities of the alternatives, and in part by e.g. a probabilistic distribution. Three examples of such models (the Dogit model, the PLC model and the Fuzzy model) are given below. Neither of these models are GEV models, rather they are models with latent preferences included.

The Dogit model

The Dogit model is one of the less well-known models described in the literature (Gaudry and Dagenais, 1978). It allows for pair wise correlated alternatives, hereby relaxing the IIA condition. The form reminds of the MNL model, though it contains an extra term that can be interpreted as an income effect, or dodged probability of choosing the hitherto chosen alternative. The model was supposed to be able to model loyal consumers as opposed to rational consumers.
3.5 Models with correlated alternatives

The model is formulated below, first the traditional formulation. Secondly, a formulation to highlight the part of the redistribution that is due to the 'dodging' or captivity of individuals to some alternative, in 3.28.

\[
P(i) = \frac{\exp V_i + \theta_i \sum_j \exp(V_j)}{(1 + \sum_j \theta_j) \sum_j \exp(V_j)} \quad (3.27)
\]

\[
P(i) = \frac{\theta_i}{1 + \sum_j \theta_j} + \frac{1}{(1 + \sum_j \theta_j) \sum_j \exp(V_j)} \exp V_i \quad (3.28)
\]

The Dogit model can handle a partial correlation structure; an alternative is dependent on the other alternatives utility by letting \( \theta_i \neq 0 \).

\[
\frac{P_i}{P_j} = \frac{\exp V_i + \theta_i \sum_k \exp V_k}{\exp V_j + \theta_j \sum_k \exp V_k} \quad (3.29)
\]

It was argued that the Dogit model could enhance modelling when different choice set applied to different individuals (Swait and Ben-Akiva, 1987), though the literature has not revealed heavy use of the model.

**The Parametrised Logit Captivity model**

The Parametrised Logit Captivity (PLC) Model is a generalisation of the Dogit model described above. The extension is of the captivity parameter which is formulated as \( \theta_i = F(\gamma Z_{iq}) \), where \( \gamma \) is a vector of parameters and \( F \) is some function to be estimated. The data \( Z \) in the captivity function is not required to be the same as the attributes \( X \). The PLC formulation is obtained by substitution of \( \theta_i \) into (3.28). The model is further described in Swait and Ben-Akiva (1987).

**The Fuzzy Logit Model**

Conventional models take as input observed or stated values of the attributes of travel. The model quality is dependent on the quality (accuracy) of the data. In fuzzy logit models an extra operation is performed prior to the estimation, which is termed a 'fuzzyfication'. This step translates an input variable into a new variable (this is termed a 'rule'). Rules may be combined using a fuzzy operator (AND or OR). Fuzzyfication can be seen as a probability model assigning a level value based on the original data, due to data vagueness or classification of data. Few papers applying the method are emerging at conferences (Holland, 2000; Mizutani, 2001).

Neither the Dogit model, the PLC model nor the Fuzzy Logit models, are GEV
models, rather they are models with latent preferences included. The class 'Latent Preference models' is emerging (Ben-Akiva, 2002), and includes the Dogit model, the PLC model and Fuzzy models.

Yet another approach is to allow for correlation between the residual terms ($\epsilon$) in the utility function, whereby models segregate themselves from those mentioned in the previous sections.

Latent Class Models

Model applications are often seen to segment the data by e.g. purpose, as years of model developing has demonstrated model improvements on this account. A simple test for whether separate models for subpopulation would improve model fit, is the likelihood ratio test of the model with a sets of coefficients for each anticipated segment versus the model with one set of coefficients.

Extending the idea of segmentation to not only being determined by exogenous variables, but merely based on a latent segregation within the data is termed latent class modelling (Walker, 2000; Walker and Ben-Akiva, 2001). Probabilistic membership functions are introduced for the latent segregation, instead of the cruel segmentation by e.g. purpose. Application of piecewise continuous distributions may be compared to fuzzy logic (see e.g. Chen (1996)) whereas continuous distributions may be modelled with the Mixed Logit model (see section 3.6.2).

3.6 Models allowing for non-deterministic behaviour

Models presented above all assume that individual behaviour is identical across the population. One simple step away from this way of thinking, is segmentation of the (sample of the) population by e.g. trip purpose, into smaller samples, consistent in behaviour within each sub-sample, and generally different in behaviour between subsamples. The idea of segmentation can be extended to one sample per individual where the difference between the samples is described by a (continuous/discrete) distribution. Such models are labelled 'Mixed Logit', 'Random Coefficient (Parameters) Logit', 'Error Components Logit', 'Models with a logit (Probit) Kernel. To the authors knowledge, the first application within transportation modelling (market shares) is (Boyd and Mellman, 1980; Cardell and Dunbar, 1980)\textsuperscript{15}. The different models are described in the following sections.

3.6.1 The Random Coefficients Logit model

Briefly formulated the construction of the utility function is altered from $U_{ji} = V_{ji} + \nu_j$, where $V_{ji} = \sum \beta X$ respective $\nu_j$ represents the explained respective

\textsuperscript{15}The work was reported previously in a technical documentation from EPRI, 1977.
the unexplained variation in the utility function for choice of alternative \( i \) by individual \( j \), to
\[
U_{ji} = V_{ji} + \eta_j + \epsilon_j. \tag{3.30}
\]
This corresponds to partitioning of the unexplained variation (the \( \nu \)) into a systematic (describable) \( \eta_j \), and an unsystematic part \( \epsilon_j \).

In Random Coefficients Logit (RCL)\(^1\) \( \eta_j \) is formulated as a mean zero random variate \( \xi \) multiplied by an attribute, potentially one for each attribute, while the unsystematic part (\( \epsilon \)) is (again) assumed to be EVI distributed - though not identical to \( \nu \). This facilitates estimation by means of the logit framework, often referred to as 'Logit Kernel'. Hence, the formulation is as follows
\[
U_{ji} = V_{ji} + \xi_j X_i + \epsilon_j = \sum_k (\beta_k + \xi_{jk}) X_{ik} + \epsilon_j. \tag{3.31}
\]
where the distribution of \( \xi \) is interpreted as the variation in preferences between individuals. The elements of the matrix \( \xi \), are either assumed to follow a stochastic distribution with mean zero and variance \( \sigma_j^2 \) or are identical 0 (fixed coefficients). Traditionally, distributions suggested are the normal (McFadden and Train, 2000) and the lognormal (Ben-Akiva et al., 1993)\(^1\) (Train, 1999), though other distributions have been suggested which include the \( \chi^2 \) (Nunes et al., 2001), the uniform (Revelt and Train, 2000), the triangular (Revelt and Train, 2000; Train, 2001; Hensher and Greene, 2003) and the Rayleigh distributions (Siikamäki and Layton, 2001). Adding a covariance term between the stochastic elements \( \xi \) is theoretically possible; though applications have been limited to the normal distribution, to the authors knowledge. The un-systematic parts \( \epsilon \) are IID.

Typically, \( \xi \) has been specified as a vector, that is, the random effects are assumed (statistically) independent, which again implies that a person with a higher than average value of travel time, is not any more likely than any other person to have a higher than average value of e.g. waiting time. Few examples of applications including \( \xi \) as a matrix (at least one non-zero off-diagonal element) are (Nielsen et al., 2001; Sørensen and Nielsen, 2001; Sørensen, 2002; Hensher and Reyes, 2000).

In the RPL formulation the coefficient referred to as \( \beta \), may differ in size from the \( \beta \) referred to in traditional (fixed coefficient model) as they are estimated from different likelihood functions. The \( \beta \) in the RPL is the mean value of a distribution which may not be unimodal, symmetrical. \( \beta \) in the traditional mode

\(^1\)In the literature the terms 'Random Coefficients Logit' (RCL) and 'Random Parameters Logit' (RPL) models are used interchangeably.

\(^2\)In this application the distributed term was the ratio of two coefficients (the VOT) rather than the coefficient itself. The idea was to reformulate the utility function such that \( U = \alpha c + \beta t + \epsilon_j = \alpha (c + \eta t) + \epsilon_j \), where distribution was allowed for \( \eta \), hereby circumventing finding the distribution of the ratio of two stochastic distributed variables.
is estimated from a likelihood function assuming identical individuals, based on a number of observations.

### 3.6.2 The Mixed Logit model

The Mixed Logit (ML)\(^{18}\) formulation is an alternative formulation of the utility function in (3.31). As the additional stochastic element(s) in the RCL are confined to the vector space given by the attributes, they may vary freely in the ML formulation. Distributed elements are added to the deterministic utility \(V\) and the aim is improved fit of the model. The RCL is a special case of the ML.

The distribution that describes the distortion between individuals is in the statistics sphere termed a 'Mixing distribution', hence the name Mixed Logit. To give an indication of the distribution of the residual term \(\epsilon\), Logit Kernel and Probit Kernel are sometimes used. A special case is the probit kernel where the mixing distribution is the normal – this again, is a probit model due to the additivity of the normal distribution.

The likelihood function to be maximised is of the form

\[
\mathcal{L} = \int_{\Theta} L(\beta|X)f(\beta|\theta)d\epsilon, \tag{3.32}
\]

where \(\Theta\) specifies the space of the distribution (if the distributions are independent the integral can be split into a sequence of one-dimensional integrals, to ease computation). Estimation of coefficients (and parameters of the distribution) is assisted by simulation, see section 4.6.1.

The literature shows an increased use of the ML model and increased understanding of the field. Examples are (McFadden and Train, 2000; Ben-Akiva et al., 1993, 2001; Hensher and Reyes, 2000). At this point it should be mentioned that, only a few statistical distributions are available in the generally available software (Alogit, SAS MLM procedure, BioGeme) or GAUSS code available on the web.

### 3.7 Models allowing for heteroscedasticity

Previously all models inherited the IID assumption, implicitly assuming that individuals act independently of one another and that deviance from average behaviour is similar for all. Heteroscedasticity is differences in the variance term and may arise as between alternatives or between observations (individuals). Between alternatives heteroscedasticity, may arise due to differences in the available level of information between often and rarely selected alternatives, the former is generally better known by the individual. Heteroscedasticity between individuals, may be caused by different types of data entering the same model.

\(^{18}\) Few papers use the term Mixed MultiNomial Logit (MMNL) for the same model.
3.7 Models allowing for heteroscedasticity

e.g. RP and SP data, where the SP alternatives are 'exact' described and perception of RP alternatives are influenced by 'non-observable' characteristics. In addition to the below mentioned models, the probit model belongs to this group due to the flexible structure of the residual terms.

3.7.1 The Heteroscedastic Extreme Value model

The terms Heteroscedastic MultiNomial Logit (HMNL) model and Heteroscedastic Extreme Value (HEV) model\(^{19}\) both occur in the literature, though the principle behind them are the same. The HEV and HMNL models deviates from the traditional MNL in the specification of the residual term \(\epsilon\). As the MNL is confined to IID, Extreme Value Type I (EV1) error terms, the HMNL assumes these independent, but not identically distributed (different variances, though shape of distribution is EV1).

For HMNL the variance specification of individuals facing the same choice situation \((E_q)\), with identical utilities (choosing differently) their scale is described as \(\mu(E_q)\). The choice probability is given by

\[
P(i) = \frac{e^{\mu(E_q)V_{i,q}}}{\sum_j e^{\mu(E_q)V_{j,q}}},
\]

(3.33)

where there are different systematic components \(V_{i,q}\) for each case \(q\). The model is consistent with RUM (Swait and Adamowicz, 1996), but is not restricted by IIA as the residual distribution varies over the population, hence the denominator does not cancel out in the ratio of probabilities. The description of the residual variance is further extended in Swait and Adamowicz (1999).

In the HEV case the variance of the error term is described by a continuous distribution as described in Bhat (1995). The specification is

\[
P(i) = \int_{-\infty}^{\infty} \prod_{j \in C, j \neq i} \Lambda \left[ \frac{V_i - V_j + \epsilon_i}{\theta_j} \right] \frac{1}{\theta_j} \lambda(\frac{\epsilon_i}{\theta_j}) d\epsilon_i,
\]

(3.34)

where \(\lambda(\cdot)\) and \(\Lambda(\cdot)\) are the density and cumulative distribution function of the EV1 distribution. \(\theta\) is the scale parameter of \(\epsilon\). The number of covariance terms to be identified\(^{20}\) in the HEV model is \(J - 1\), which is a substantial reduction compared to the Probit model that includes (up to) \(J(J - 1)/2\) covariance terms. Nevertheless, estimation becomes slightly more complicated as the HEV is not on closed form. A suggested method to overcome this, is by application of Gaussian quadrature or simulated maximum likelihood (pseudo-random or Halton draws), described in chapter 4.

\(^{19}\)The model is also referred to as Heteroscedastic Extreme Value Logit, HEVL

\(^{20}\)Identification of models is discussed in section 2.3.2.
3.7.2 The Heterogeneous Conditional Logit model

The Heterogeneous Conditional Logit (HCL) model (Steckel and Vanhonacker, 1988), can handle continuous heteroscedasticity between observations (individuals). The model is, as most other models, based on the MNL model (assuming EVI residual terms) but adds the individual heterogeneity by including a specific scale parameters for each individual. The model can be shown to be a special case of the ML model. In (Steckel and Vanhonacker, 1988) the scale is assumed gamma distributed.

3.8 Elasticities

The area of demand modelling is not confined to the above mentioned methods. Other techniques are applied on their own or in conjunction with above mentioned methods. Elasticities are a tool to consult, when the expected change in probability given some specified change (in some variable) is needed. Though, the estimate is rough (only valid for very small changes, as it involves partial derivatives) it is frequently used due to the almost zero cost of obtaining it.

Rather than providing a huge table of not immediately comparable elasticities for different models, the general formulation is provided.\textsuperscript{21} This formulation also applies for Mixed Logit/Hybrid choice formulations, though the partial derivatives are harder to obtain. The elasticity for the probability of choice of an alternative with respect to a change in an attribute relating to the alternative in question (i), is defined

\[ E(P(i), x_{ik}) = \frac{\partial P(i)}{\partial x_{ik}} \frac{x_{ik}}{P(i)} = \frac{\partial \ln P(i)}{\partial \ln x_{ik}}, \]  

(3.35)

whereas the elasticity with respect to a change in another alternative’s attribute is given by

\[ E(P(i), x_{jk}) = \frac{\partial P(i)}{\partial x_{jk}} \frac{x_{jk}}{P(i)} = \frac{\partial \ln P(i)}{\partial \ln x_{jk}}. \]  

(3.36)

3.9 Discussion

The chapter demonstrates that quite numerous (variants) of choice models exist. The listing was intended to give a grasp of differences between and limitations of the models. Developments within modelling, is linked to developments within computer power, as many of the models emerging today were not possible to estimate 5-10 years ago on an affordable computer.

\textsuperscript{21}When models are compared by elasticities more principal differences (e.g. distribution/correlation in residual term and other model assumptions) are often left unaccounted for. The direct and cross elasticities are written out in the papers describing the models, see the references for the relevant model.
3.9 Discussion

In the past decade the evolution of models seems to have been twofold: description of similarities by means of rigorous nesting structures, along with developments within models using stochastic elements to a wider extent. In both cases, more expertise is demanded of the researcher/practitioner to be able to correctly specify and interpret results of models. This, in turn, requires a larger budget for the project or recognition of the level of accuracy required for the given project. Whether new model variants are used or not, awareness of the models capabilities is increased.

Models presented in the previous sections each and one, and in different ways, add on to the logit model (apart from the MNP/Probit Kernel models). Segmentation, distribution of coefficients, heteroscedasticity of residual terms and complex nesting structures to enable non-independence between alternatives. As the list of models is quite long, the models are summarised in figure 3.1, starting with the simplest model at the lower level and complexity rising upwards. Arrows point towards model of greater generality – the most general at the upper level. The three different model classes are given (from left) as Hybrid Choice Models, Tree Models and Latent Class Models. The red dotted line indicates that the ML can approximate any GEV model. Models names in italic refer to models where the residual term is Normal distributed, otherwise the residual is distributed EV1.

The RNEV (and a number of it’s special cases; CNL, PCL, NL, GNL and GenMNL) involves tree structures with quite numerous different structures. A critical issue of using these models is whether the best tree structure is found. A complete search over the different structures may lead to problems of Type I and II (rejecting (accepting) a true (false) hypothesis). Another approach to follow is the employ the Paired version of the GNL model to search for nesting structures/similarities and thereby reducing the number of tree structures substantially (Wen and Koppelman, 2001).

Recent research has shown that models of complex nesting structures and models of complex stochastic structures are similar as any tree structure may be approximated by a ML specification (McFadden and Train, 2000). Yet it is realised, that the number of different nesting structures (no restrictions on alternatives) rises extremely fast with number of alternatives involved. Allowing for alternatives to appear in more than one nest (or to be chosen via different paths through the tree) only worsen this problem. As for the mixed models the issue of constructing the models is either diverted to obtain a good model fit with a few additional terms, or to gain understanding of what drives the individual’s choice process (tastes), which may require a (larger) number of terms possibly correlated.

Future research will, I believe, improve the understanding of the possibilities and limitations of the mixed models and models of non-independent alterna-
Figure 3.1: Relations between discrete choice models
3.9 Discussion

tives. As understanding of the models have remarkably improved, new gaps have been exposed; 'comprehensible' models (cross-nested or 'simple' tree structure), aim at model fit or understanding of individuals choice formation, how many distributed terms to employ in an model, which distribution(s) to apply, and whether the distributions should be correlated?
Chapter 4

Data and estimation

Specification of model form, that is, whether a logit model or a mixed formulation should be applied is a good beginning, yet there is a long way to go. Model coefficients must be estimated, which in turn requires the existence of estimators for each of these. Further, it requires some data in order to estimate the model coefficients. This chapter addresses these issue, from data sources, differences between types of data and potential bias in these, over existence of estimators to approximation techniques when estimators are not readily available. Three different types of estimation techniques are described; namely closed form estimation, simulation assisted estimation and sample based estimation.

4.1 Data

Estimation of a model is subject to the data it is based on. Further, the quality of the model is subject to the quality of the data it is based on. Therefore, improving the estimation methods without looking at the quality of data is like wetting one’s pants to keep warm! Data quality is described in section 4.2. First a description of terms related to data is provided - such a description can be found in most introductory statistics books, e.g. Glenberg (1988); Thyregod (1993) or books dedicated to sampling theory e.g. Cochran (1977).

Starting with the simplest level, the observed value for each of the variables included in the survey. These values are defined as

**Measurement** is the use of a rule to assign a score to a specific observation of a variable.

The score is also referred to as the values of the explanatory variable. Further, a characterisation of whom or what the information stored in the scores relates to, is needed. This unit of reference is described as;
Measurements are assigned given some choice of sampling unit. Units are non-overlapping and exhaustive of the population.

In relation to surveying transport, the choice of unit can be individuals, households, two-worker households, or at more aggregate level, traffic between zones, or some other pre-specified unit. Surveys can be designed to capture the travel patterns of these units. In this thesis, the unit is ‘trips’. Conducting travel surveys and/or traffic modelling, each record is often referred to as ‘an individual performing a trip’ or just ‘an individual’. Rather, the individual is the score of the variable containing information on the persons identity.

The collection of scores for explanatory variables, related to one discrete choice is often termed a ‘record’ or an ‘observation’. Formally,

An observation (record of scores) relates to one choice undertaken by one sampling unit.

An observation is the collection of scores (one for each explanatory variable plus the variable containing the choice) for one choice situation. In the thesis ‘observation’ and ‘record’ are used interchangeably.

One individual may, or may not, have caused more than one observation in the data set. This is captured in the definition of the population:

A population is a finite or infinite collection of 'units'.

In a population consisting of all trips performed on a specific Sunday, some individual may give rise to two observations in the population, if that individual performed two trips on that specific Sunday. The population we are interested in, is the statistical population. A statistical population does not necessarily consists of people; rather it is a set of scores (measurements of a variable) that can be related to an individual or some other entity we are interested in. The common characteristic describing a population may be ‘all residents of some particular area’, ‘all individuals performing a shopping trip on Saturday’, ‘all cars parked on street at weekends’, etc.

A population can be formalised as \( \{ x(i) \}_{i \in I} \), where \( I \) is an index set (\( I \) may be structured as \( I = 1, \ldots, N \), but this is not strictly required). If the population is finite, the index set \( I \) is finite, vice versa. The scores are functions \( x : I \rightarrow X \), where \( X \) is the space of possible values for the variable. The scores are \( x(1), x(2), \ldots, x(N) \); these may describe e.g. number of trips performed between 8am-10am, household income, etc.

Say, the population we are interested in consists of individuals performing a trip by any mode, this population may prove quite burdensome (and expensive) to collect data for, simply due to the size of it. For this reason, a subset
4.1 Data

of the data is often used to infer about the population, where sub-set reflects that fewer units than the whole population are included but also that fewer attributes than those characterising each unit, are included.

A sample is a subset of scores from a population.

A sample of size \( n \) from a population \( \{x(i)\}_{i \in \mathcal{I}} \), is a set \( s = (i_1, i_2, \ldots, i_n) \) from the index set \( \mathcal{I} \). The scores in the sample are \( x(i_1), x(i_2), \ldots, x(i_n) \).

A random sample is selected so that every score in the population has an equal chance of being included.

Looking solely at the sample, it cannot be determined whether the sample is a random sample or 'just' a sample. But the impact on later analysis is important. If the scores in a (non-random) sample were selected due to their relatively high probability of being included, this should be reflected at later stages of the analysis – otherwise conclusions are likely to be erroneous. In practice and if possible, random samples are often aimed at as these enables further analysis without having to correct for un-equal probability of being chosen.

A number of different techniques are employed for constructing random samples, these are described in section 4.4. Below, the nomenclature related to selecting samples and reference to sub-populations are given.

A strata is a sub-set of the population. Strata’s are non-overlapping and totally exhaustive of the population.

Assessed information on the part of stratas which are in some specific sample, can be enumerated to the population, given that the size of each strata in the sample and in the population are known. Within transport modelling, a division of a population of trips, by some trip characteristic is often convenient. This is generally referred to as segmentation, though it is in fact a stratification. The most common stratification of data is by trip purpose, though segmentation could be by other trip characteristics. Segmentation is further described in section 4.3.

Two final terms used to refer to sub-sets of the data sample are 'group' and 'cluster'. In this thesis 'group' refers to subsets of the sample which may be fairly equal in size size (division of 20 units into three groups cannot be identical in size) or may differ in size.

A group is a sub-set of a data sample. Groups are non-overlapping and the collection of groups is exhaustive of the sample.
Data groups can be constructed by simple division of data proportional to one (some) variables, random sampling, cluster analysis etc. A cluster on the other hand, stores some information of the 'selection criterion' within the cluster as,

\[
\text{A cluster is a sub-set of the data sample. The sampling units are grouped into clusters sharing some common feature.}
\]

Clusters do generally vary in size, the collection of clusters will be exhaustive of the sample, though some of the cluster may consist of just one unit. Clustering procedures are described in section 4.4.2.

### 4.2 Data quality

Quality of data is a queer specimen, in the way that it consists of several characteristics. A large data sample, many attributes of travel, socioeconomic characteristics of the individual and household, multiple realised or stated choices for each individual, etc. are all nice properties of a data set. 'High quality data' is data that meet all these demands, whereas 'low quality data' may fail on one or more of the issues. Unfortunately, high quality data has a price – data collection is expensive and increases by number of observations, number of variables included and increasing accuracy of the recorded answers.

Different precautions must be taken based on how the data collections is actually undertaken, that is, whether or not an interviewer is present while the respondent answers the questions (or the interview is performed by phone), whether the questionnaire is printed or provided on pc/ by internet. An interviewer can aid in clarifying the questions/answer options to the respondent and thereby reduce the number of 'don’t know' answers and unanswered questions. On the other hand, an interviewer may cause bias by the way the questions are read out loud. Generally, use of an interviewer improves the data quality.

Presenting the questionnaire on paper often makes the respondent feel confident, as most people have seen a paper based questionnaire before. Furthermore, it has the advantage that the respondent is self-sufficient, the questionnaires can be mailed to respondents and/ or mailed back, which all reduces cost of the survey. From the researchers point of view, the pre-printed version is rather rigid; if a respondent does not have the option of altering the time of travel, any question related to time of travel is irrelevant for this particular respondent – but certainly relevant in relation to a respondent that has the option.

The counterpart, a dynamic questionnaire that adapts to the answers given by the respondent, is performed by use of a computer. It gives the researcher the opportunity to close in on the trade off between specific attributes of travel, e.g. how many minutes of additional waiting time corresponds to cost reduced by 5 Dkr? Model coefficients are subsequently estimated from the respondents
4.2 Data quality

stated trade off between attributes, where the questions are designed to narrow the trade off between attribute values for changed choice of alternative. The interview type is restricted by the need for an interviewer (to guarantee the respondent answers consistently with the question and answer options, reduce number of typing error and ensure that the respondent does not loose overview of the questionnaire and skip).

Several sources of bias have been identified in the literature (see e.g. Ben-Akiva and Lerman (1985); Richardson et al. (1995)), and are listed below:

1. Lexicographic choices
2. Deliberate respondent bias for political reasons
3. Respondent fatigue
4. Misunderstanding of the way questions are posed (by one respondent or all interviews performed by an interviewer)
5. Systematic neglect/omittance of an alternative
6. Lack of 'don’t know' or 'no choice' answer option
7. Additional, non-reported information considered by the respondent.

The first two sources of bias (choice based solely on one characteristic of travel and respondents trying to affect the outcome of the analysis), are performed (deliberately) by the respondent. Sources 3 and 4 are caused by not carefully planning the questionnaire, that is, a bias caused by the researcher. These four sources of bias are to some extent identifiable in the data, and may therefore be accounted for by leaving the question, the record or the observation, or even all interviews performed by one interviewer, out of the survey data in the model estimation.

The fifth source arises if a respondent for some reason, does not wish to consider a specific alternative, e.g. the person will never consider using the bus, even if it was for free! A respondent may not report choice of some alternative, but the reason may be that other alternatives are more attractive – and not an aversion for the mode. Therefore, such a bias may not be directly identifiable from data.

---

1The term ‘Lexicographic choices’ covers (at least) three different issues. First, the case where the individual always chooses say car, irrespective of how the alternatives are described (an alternative mode may be free of charge). Second, the choice is dominated by one characteristic, say travel time, where the alternative with the lowest travel time is chosen, no matter which type of mode it is and any other characteristics. Finally, the data reported on some individual may seem as if it resulted from either of the two causes just described (either one mode is preferred in all situations, or, for all the choice situations, the chosen alternative has the highest/lowest value for one of the characteristic). Though, in fact the choices are the result of that particular individuals careful consideration, given some trade-off between alternatives. The third type is difficult to separate in data, but may be avoided by carefully planning.
The sixth type of bias arises when a respondent is forced to make a choice of an alternative, when neither of the alternatives are preferred. An example is when two different modes are described for a shopping trip, and the respondent will not perform the trip, given the information. Or, if description of alternatives is closely related to a real choice situation, the individual may know a better alternative. This is only relevant when the questionnaire cannot be continued without an answer (e.g. pc based, non-assisted questionnaire). The last type, is when the respondent uses some prior information on (one or all) modes in a choice between alternatives. Information like 'busses are usually late', may affect the choice but this is not reported with the data. In the model estimation the choice of alternative is attempted explained by the available information (the attributes describing the alternative). As not all information used in determining the choice is available, the variance on the residual term increases causing a poorer performing model.

In the list above simple misunderstanding between respondent and interviewer should be included. Explained in more detail, the researcher is interested in information on some issue and decides to form a question. This question may or may not, cover exactly what the researcher had in mind, but if the respondent does not understand it exactly the way the researcher had in mind, it is not good. Then, the respondent determines an answer and writes/ states it, and again if the researcher does not understand it exactly the way the respondent had in mind, it is not good. This leaves us with four chances of distorting the data, even if everyone did their best to collect the data. This is illustrated in figure 4.1 where the arrow is the processing of the questionnaire.

Even for carefully validated data, careless handling and interpretation of data may cause problems in modelling. A classical error is that when several data sources are joined, several travel time variables arises. Say, one variable is door-to-door travel time and a second variable is in-vehicle travel time. These variables should be treated as two different variables in the modelling, but if there is incomplete information about what kind of travel time is stored in the variable labelled 'travel time' this is not an option.

### 4.2.1 Data types

The most common means of data collection are either to observe or state travel patterns for some number of persons. Observed traffic is in general referred to as Revealed Preference data, while information on individuals’ presumed actions (given some scenario) is named Stated Preference.

**Revealed Preference** (RP) data (traffic counts) are easily accessible for corridors like ferries, tolled roads/bridges and air traffic, whereas counts of traffic between e.g. two major cities involve counts on each of the connecting road (possible deflated for through traffic). Collections of RP (time series) are used
4.2 Data quality

What the researcher *thinks...*  ..the respondent answered...

...the researcher posed.  ...to the question the respondent *thought...*

Figure 4.1: Bias in reporting data

for studying the development in traffic patterns; extension, distribution and total volume.

**Stated Preference** (SP) data can be collected by observing the 'choice' of a respondent, when presented for two (or more) artificial alternatives for a specified trip. If data collection is performed by pc, the questions can be tailored to the respondent in the sense, that the alternatives relates to a trip the respondent is familiar with e.g the home-to-work trip and levels of variation for each attribute can (in principle) be narrowed in to more adequately depict the (relative) preferences.

**Rank order** data requires the respondent to order alternatives by their preference. Information is given on the order of the preferred alternatives, though not on magnitude of the difference. The related, though different, **rating order** is grading of the alternatives. Here two or more, alternatives may obtain the same rate, which reflects indifference between the alternatives.

Previously used data collection methods include **Transfer price** (TP) and **Stated Intentions**, where TP may be about to experience a renaissance.

Different data sources are used in conjunction when building a model using
the approach of a single element nested logit model (Daly, 1987). The basic idea is a nested logit model where the pseudo alternatives just below the root are (alternatives subject to each of) the data sources. The scaling introduced between the data sources accounts for potential differences between the sources (different respondents, different time periods of interview, different geography etc.).

The design of a questionnaire should at any time, keep in mind the model developing to follow as the model cannot test the effect of a variable, if the design does not allow for such a specification. An example is a questionnaire (whether full factorial or fractional factorial) with perfectly orthogonal design, which does not allow for testing non-linear synergy between variables. Similarly, if a model should be able to consider both small changes and large changes in the same variable (e.g. travel time) the design of the questionnaire must reflect this.

4.3 Data Segmentation

Formulation of a model on some collected data involves the choice of probability model but also taking a stand on whether it is likely that the whole data set can be described by the same utility function. If it is not believed that one utility function can describe the variation in the data, more utility functions can be formed where each of these applies to (describe) the variation in a part of data or data segment. The collection of utility functions describe the whole data set. Further, as the data set is a (representative) sample of the population, the population can be described by the collection of models for data segments.

A explanation of what segmentation is, could read like

*The division of an image into regions or objects. Often a necessary step before any quantitative model can be carried out.* (Everitt, 1998).

To be more concise and adapt the formulation to transport modelling, the objective is to partition a data set according to some rule, such that all elements of a data set is unambiguously assigned to exactly one segment.

This section describes the ‘common practice’ on data segmentation and sampling, where issues are described and the literature presented.

The general use of the term *segmentation*, refers to models where sub-models are constructed for segments formed on the basis of a simple criterion. Probably the most widespread use is to segment by trip purpose (home-work, business, shopping, leisure, etc.), by income level (low, middle, high). Other used standards are car availability for the present trip (no, limited, full) or activity type (worker, non-worker). A similar term *latent class* (Walker, 2000; Ben-Akiva
4.3 Data Segmentation

et al., 2001) also refers to formulation of different utility functions for sub-sets of data. This is described further later in this section. Within statistics research, the term strata is used for sub-samples of a data set, where the members of each strata is chosen such that they are characterised by certain attributes of interest. In essence, segmentation is stratification.

In the literature on passenger transport modelling, no unambiguous definition of data segmentation exist; though quite a number of models reported did include data segmentation. Examples of how the use of segmentation in models has been reported are ...results were obtained for each of six market segments (commuters, business travellers, education and other personal travel purposes, light and heavy goods vehicles...) (Nielsen et al., 2000) and ...All the parameters were estimated for segments which are different in activity type and worker/non-worker... (Arentze et al., 2000). Further, the choice of a particular segmentation may not be validated in the study, but may be motivated by previous studies ..these categories are selected because previous research indicates that the age and employment status captures a significant portion of the variance in activity-travel behavior (Kulkarni and McNally, 2000). In a study where segments (latent classes) were compared to an unsegmented model it was concluded that 'the 2 latent classes do provide a significant improvement in fit over the base [unsegmented] model' (Walker, 2000), rather than determining whether the within-segment variation was significantly lower than the within-data variation.

A number of larger national traffic models include segmentation. Examples are the Danish National Traffic model (Landtrafikmodellen) (Larsen and Filges, 1996), the Swedish National model (Algers and Gaudry, 2002) and the Öresund model (Pedersen, 1998; Sørensen et al., 2002). Further a Swedish value of time study applied different segmentation criterions to the same data in Dillén and Algers (1998).

From this point the following formulation of segments and models based on segments are applied.

A data segment is a subset of a sample, which is unambiguously described on the basis of a (few) simple criterion(s).

A segmented model is the complex of a data sample, where separate models are constructed for each data segment.

4.3.1 Mathematically

Data segmentation is the division of a sample of data into a number of sub-samples. The purpose of segmentation can be twofold: to get an improved description of sub-samples of data and hereby data in general, and to be able to im-
prove the model forecast, by improved forecasts for sub-samples of data and the extrapolation of these sub-samples over time. Formally, a data set $X$ (the parent data set) may be divided into segments $X^1, X^2, \ldots, X^m, \ldots, X^M$ where sub-samples are nonempty $X^m \neq \emptyset, \forall m$, non-overlapping $X^m \cap X^k = \emptyset, \forall m \neq k$ and totally exhaustive $\bigcup_m X^m = X$.

The choice probabilities for choice of alternative $i$ within the parent data set $X$ are described by

$$P(i|\beta; X),$$

(4.1)

where $\beta$ is the parametrisation of the model. Estimation of coefficient values requires the assumption of probability model e.g. MNL (see section 3) and the use of an optimisation techniques e.g. Maximum likelihood (see section 4).

The parallel formulation for for choice of alternative $i$ for individuals in data segment $m$, can be described by

$$P(i|\beta^m; X^m),$$

(4.2)

where $X^m$ is the data describing the segment. As indicated by the parametrisation $\beta^m$, the parametrisation for segment $m$, may divert from the parametrisation $\beta$ for the parent data set and that for another segment. Differences between parametrisations are due to differences between the utility function; e.g. one variable (and thereby the coefficient associated with it) is only included in the utility function for one segment.

The probability for choice of alternative $i$, based on the data segmentation $X^1, \ldots, X^M$, can be described by

$$P(i|\beta^1, \ldots, \beta^M; X^1, \ldots, X^M) = \sum_m P(m)P(i|\beta^m; X^m),$$

(4.3)

where the choice probability for each alternative, for each segment is weighted by the probability for membership of that segment $P(m)$.

Often, the construction of a model has two purposes, namely the descriptive and the predictive (forecasting) purposes. The descriptive part of the model formulation enables the researcher to get an indebt understanding of each segments behaviour and further the differences between segments. Enumeration of the sample to the population, can be used as a validation tool – to justify whether the estimated model can explain the (aggregate) choices of the population. The enumeration requires the relative size of the segments is known for the sample as for the population. The need for knowing the relative size of the segment in the population may restrict the list of potential segmentation criterions, simply because the relative size in the population is not known nor can be obtained.

In the case of forecasting, $P(m)$ will reflect the expected probability for membership of the segment within the population, at the forecast time. If different
4.3 Data Segmentation

models are estimated for different data segments and the composition of the population is expected to change over time, the researcher is enabled to incorporate these changes by use of a segmented model. The alternative, using the same model for the whole data would leave the researcher without a possibility of incorporating this information.

Forecasting in a segmented model is discussed in section 4.8.

Existence

In order to apply data segmentation in model, at least a comment on existence of segments is needed. Recall, that segmentation is a way of splitting a data set into subsets by some predefined criterion. The issue of existence of segments within a data set \( X = \{x_1, x_2, \ldots \} \) is trivial, unless for the cases where A) the data set is empty, B) the variable defining the segment membership is either undefined for all elements in the data, or C) they all belong to the same segment (records are similar). Neither of these cases are particularly interesting, hence, existence of segments is assumed for the rest of this thesis.

Latent classes

Latent classes are designed to capture unobserved heterogeneity generated by discrete constructs, which are not directly observable and therefore represented by latent classes. In the formulation of the latent class model in Ben-Akiva and Bierlaire (1999), it can further handle that different choice sets \( (C_1, \ldots, C_M, C_m \subseteq C, \forall m) \) are available for the segments. The principle is similar to the principle of segmentation described above, where the probability for choice of alternative \( i \) within each class \( m \) is weighted by the probability for the class \( P(m|X_n, \theta) \), also referred to as the ‘class membership model’. The formulation for the latent class model is given below.

\[
P(i|X_n) = \sum_{m=1}^{M} P(i|X_n, \beta_m, C_m)P(m|X_n, \theta) \tag{4.4}
\]

where \( C_m \) is the set of alternatives and \( \theta \) is the unknown parameter vector. \( X_n \) refers to the data for individual \( n \).

Concerning the broader definition of latent classes, membership of a class may be determined by a probability function. In this case it may not be evident which class each data record is a member of. Though as long as the sum over probabilities for membership of the classes \( P(m) \) is 1, whereby the probability for not being assigned to a class is 0, the expectation is that classes do exist.

4.3.2 ‘Hierarchical’ segments

Segmentation is a division of data into smaller sub-samples (segments). If the procedure is repeated for the segments such that sub-segments are appointed
within each segment, 'hierarchical' segments are obtained.

The motivation for performing the segmentation procedure is to first perform a coarse segmentation to enable detection of large differences. The subsequent formation of sub-segments can be based on a more sensitive procedure. Further, if a sample is constituted by e.g. 15 segments, it may be more comprehensible if presented as 'three segments each with five sub-segments'.

The detection of segments can be performed either as 'top-down' or 'bottom-up'. In the 'top-down', a few general segments are pinpointed and a search for sub-segments is conducted for each of the segments. Hereby, the description of sub-segments may vary significantly between segments. In the 'bottom-up' a range of segments (the sub-segments) are found which are subsequently united into larger groups (the segments). This way, some consistency between how the sub-segments are formed is ensured, as they are all identified at the same time.

An example of a hierarchical segmentation of a population is shown in figure 4.2, where the first segmentation is by number of cars in the household (HH) illustrated by dotted red line, and subsequently by the trip purpose.

4.4 Sampling techniques

Usually when any kind of model of some real world issues is about to be constructed, data foundation must be considered. Making a model based on the whole populations is not a problem – at least not if the population is kids aged
4.4 Sampling techniques

10-12 years, owning a bike, living on a specific street in some particular city. However, in most cases where a model of the travel behaviour is requested, the population is a great deal larger and the members are less easily identifiable (which of the persons riding a regional train live in some specific geographic area?).

Collecting data for a large population can be very expensive. In the 'cheap' end of data collection methods is searching for any available data, where information may be aggregated to some level (trips between stations/towns, by train and bus; count of cars on a specific road segment). This information is very useful for calculating the total number of trips performed between local areas (an OD-matrix), whereas it is less useful for assessing the behaviour of the individuals who actually performed the trips.

The expensive option is to interview 'anybody' who is a potential member of the population (to ensure everybody in the population is interviewed) on all issues related to their travel behaviour. In between of these two extremes is the option of interviewing a sample of the population. The sample is selected such that it is representative of the population, in the sense that it is possible to infer about the population from the sample. Budget constraints for the survey and requirements for the accuracy of the estimators then determine the minimum size of the sample needed. The variation of the estimators and the relative size of the sample are inversely related (see Cochran (1977)).

4.4.1 Random sampling

The commonly used approach to select a sample from a population is by use of random sampling, though other types of sampling can be used. The motivation for the requirement of a random sample is that if the sample is random, the sampling distribution will be the normal distribution with the 'true' mean of the population, the distribution of the \( t \)-statistic will be \( t \) distributed and finally, the well-established inferential procedures will work. If the sample is not random, a bias is introduced which causes a statistical sampling or testing error by systematically favoring some observations over others. A random sample can be described;

\[
\text{Each observation (unit) in the population, has an equal chance of occurring in a random sample.}
\]

Random samples may be generated by making a list of all members of the population and use random numbers to determine whether each and one of the units are in the sample or not. The draw of members for the random sample is performed with replacement such that the probability for drawing a specific unit is unaffected by how many units have been selected before. Otherwise, for a population with \( N \) units, the probability for selecting each unit would increase for each chosen unit, that is probabilities would be \( 1/N \), \( 1/(N-1) \), \( 1/(N-2) \)
CHAPTER 4. DATA AND ESTIMATION

etc. With replacement, the probabilities are unaffected by the history of draws; \(1/N, 1/N, 1/N\).

To obtain a pre-specified level of accuracy of the estimates the relative size of the random sample decreases with the size of the population.

**Stratified random sampling**

A variant of the random sample is the stratified random sample, where the population is separated into sub-populations (stratas). The stratas are exclusive and collective of the population. A random sample is drawn from each strata and the collection of these random samples is the *stratified random sample*. The technique ensures representation of each of the stratas in the drawn sample at the cost of surveying more units and it is more time consuming.

Sampling techniques are further described in Cochran (1977).

**4.4.2 Cluster analysis**

Cluster analysis is a statistical tool used to find groups of observations in a data set, where the groups are characterised similarly. That is, cluster analysis can be used to detect areas with a high density of observations (a cluster) in a multidimensional space. Unlike the random sampling a great deal of information is stored in these clusters, as they have similar values for a (function of) variable(s).

Cluster analysis is a tool for grouping data by the density distribution of one/some functions of variables; though units within the same cluster may be just as different in other variables as the units in the sample are, in general. The number of clusters is not limited (no more clusters than there are data units), the size of each cluster is limited by one and the total number of units, just as there is no restriction on that clusters are ‘fairly equal’ in size. The dimensionality of the space of the variables used for the clustering, is not restricted (not infinite). Having detected these dense areas, clusters are formed around them such that each point (data record) is assigned to the ‘closest’ cluster; all points are assigned to exactly one cluster.

The idea is that clusters are formed such that a measure of between clusters difference is maximised (e.g. between clusters variance in a (some) variable(s) is maximised) or, with-in cluster distance is minimised. Depending on which criterion is applied for the clustering, cluster of different shapes are created (circular, elongated, very dense with many small clusters – which primarily consists of outliers). Clusters may be constructed in several ways; examples are hierarchical clusters (data sequentially split into disjoint clusters) or clustering with a fixed number of clusters (data is divided into groups and observations are ‘moved around’ until the clustering is the most compact).
The actual construction of clusters is based on specified variables (which should be no less than 2) on which, the resulting clusters depends, that is, clusters may change if variables used for their formation are changed/replaced. The measure of 'distance' between clusters can be the (Euclidean) distance between observations, distance between cluster centroid’s, likelihood of clustering etc.\(^2\) However, the actual choice of measure of distance may influence the resulting clusters and may thereby affect the results of the traffic modelling.

There is no criterion for choosing the best cluster procedure for a general problem - some prior knowledge of the data is required, e.g. of the shape of clusters that is likely to appear. One approach could be a combination of 2-5 different procedures where only the common patterns of the various cluster structures are applied. The cluster methods does not guarantee on finding the overall best cluster structure; but just the best clusters for a given number of clusters (a maximal number of clusters is specified and the best cluster structure with no more than the specified number of clusters is found). Therefore various upper bounds should be tested; though keeping in mind that the total number of coefficients is proportional to the number of clusters.

\(^2\)All major statistical packages can perform cluster analysis, and generally, they include quite a number of different distance measures.
Cluster criterions

A number of different approaches are applied for cluster analysis in the literature - and unfortunately no (all purpose) superior method has been discovered. The main controversy with these methods are that they are heuristics rather than analytical methods - wherefore strict convergence to the overall solution (the 'true' data partitioning) is not guaranteed. Depending on the 'true' shape of clusters in the data (well-separated, poorly separated or elongated clusters), different methods appears to be best at correctly finding these clusters. Further, most methods are biased toward finding clusters possessing certain characteristics related to size (number of members), shape, or dispersion. Most of the cluster criterions are hierarchical techniques – examples are disjoint clusters, hierarchical clusters, overlapping clusters. An example of a non-hierarchical cluster criterion is the $K$-means cluster (it forms the $K$ most distinct clusters). The measure of 'most different' is based on a measure of distance.

Distance Measures

The measure of 'most different' is based on a measure of distance or dissimilarity between objects, between clusters or between an object and clusters. When the clusters analysis initiates, the relevant distance is between objects (no clusters are formed), but as some objects are grouped into clusters the distance from an object to a cluster becomes relevant, eventually, the distance between two clusters becomes relevant. First distance measures between objects is described, then follows measures of distance between/ to clusters.

Measures of distance between two objects can be computed based on a single dimension or multiple dimensions. The simplest measure of distance between objects in a multi-dimensional space is to compute Euclidean distances $dist(x, y) = (\sum_i (x_i - y_i)^2)$, though this has the disadvantages of being insensitive to addition of new observations and sensitive to the scale of the dimensions (this may alter the clusters formed). Alternative measures of distance are City-block (Manhattan) distance, computed as $dist(x, y) = \sum_i |x_i - y_i|$. Results are similar to those for the Euclidean, though the effect of outliers is dampened. A distance measure used if difference in any dimension is the criterion for defining objects as different, the Chebychev distance given by $dist(x, y) = \max_i |x_i - y_i|$, could be used. Or the Power distance given by $dist(x, y) = (\sum_i |x_i - y_i|^p)^{1/r}$, where $r, p$ are user-defined parameters.

An amalgamation (or linkage) rule is used to determine which cluster an object is closest to. This is used to determine when two clusters (or an object and a cluster) are sufficiently close to be linked together. The essential theme is to define to which point within a cluster, the distance is measured. The distance can be measured to the closest neighbour or single linkage whereby the resulting clusters tend to be "stringy".
4.5 Closed form estimation

The opposite, to define the distance between clusters as the maximum distance between any one object in one cluster and any one object in the other cluster, is furthest neighbour or complete linkage. If the clusters are naturally distinct this method performs well, whereas elongated clusters may cause trouble.

In the **Unweighted pair-group average** method the distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters. This is also very efficient when the objects form natural distinct "bundles", however, it performs equally well with elongated clusters. There is a weighted variant where the number of objects in a cluster is used as the weight.

In the **Unweighted pair-group centroid** method, the distance between two clusters is determined as the difference between the centroids, where the centroid is the average point in the multidimensional space defined by the dimensions. Hereby, the method is less affected by the outermost objects, though the spread within the cluster is not accounted for. Again, there is a weighted variant, where the number of objects in a cluster is used as the weight.

A final method worth mentioning is the **Ward’s** method. This method is distinct from all other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In short, this method attempts to minimize the Sum of Squares of any two (hypothetical) clusters that can be formed at each step. In general, this method is regarded as very efficient, however, it tends to create clusters of small size. Further details of this method are given in Ward (1963).

Variables applied for the cluster creation should be chosen with care, as the clusters generated are subject to these. Discussions of cluster procedures can be found in a variety of statistical handbooks, see e.g. (Kotz and Johnson, 1982-99) and applications within applied research areas like e.g. market analysis.

### 4.5 Closed form estimation

Discrete choice models on closed form, may be estimated by any search algorithm or iterative solution method. Briefly described, any iterative solution method, involves a starting point $\beta_0$, a step length $\lambda$, a step size $M_t$ and a direction, often given by the gradient of the objective function $g_t$. From the starting point, the coefficient changes as given by the step

$$
\beta_{t+1} = \beta_t + \lambda M_t^{-1} g_t,
$$

(4.5)

where different specifications of $M_t$ gives rise to different algorithms.
Newton-Raphson

The most traditional approach, and probably the first choice of algorithm for most optimisation problems, is the Newton-Raphson (NR) algorithm. In the NR, the direction is the gradient $g_t$ and step size $M_t = -H^{-1}$, is the negative, inverse Hessian.

As the derivation of the iteration formulae involves a second order Taylor approximation, the method works best for problems with a formulation close to quadratic. Similarly, the steepest ascent (or steepest descent, depending on the context) is based on only first order Taylor approximation, whereby the direction is based solely on the gradient and not the curvature in the point $\beta_t$.

The step length is set to 1 in early versions of the NR, whereas letting the step length decreases by e.g. $1, 1/2, 1/4$ as long as $\mathcal{L}(\lambda_{k+1} \beta_{t+1}) > \mathcal{L}(\lambda_k \beta_{t+1})$ reduces the calculation time, vice versa.

BHHH

The Berndt, Hall, Hall & Hausman (BHHH) algorithm is based on equation (4.5), where the step size is altered compared to the NR algorithm. The step size $M_t$ applied is the average outer product of scores $s_n(\beta_t)$ in the sample, given by:

$$M_t = \frac{1}{N} \sum_n s_n(\beta_t) s_n(\beta_t)'$$

This formulation is an approximation of the true curvature given by Hessian matrix. Using this approximation rather than the true curvature, reduces the computation time as second derivatives are not evaluated at each step. Further, the BHHH is guaranteed to improve the likelihood value at each iteration, though steps can be small far from the maximum (Train, 2003).

A variant of the BHHH (BHHH-2), correcting for average scores not being 0 for $\beta \neq \beta^*$, is found by letting $M_t = \frac{1}{N} \sum_n (s_n(\beta_t) - g_t)(s_n(\beta_t) - g_t)'$. The BHHH algorithm was originally suggested in Berndt et al. (1974) and is further described with the BHHH-2 variant in Train (2003).

A drawback of the NR and BHHH methods, is that they only work for concave functions (or convex functions, by a sign change). If the function involves larger areas of local non-concavity (non-convexity), more special algorithms must be employed.

Other algorithms are the DFP (Davidon-Fletcher-Powell), BFGS (Broyden-Fletcher-Goldfarb-Shanno) both based on arc Hessian, simulation based (Monte carlo) methods like the Gibbs sampler or Simulated annealing. Quite a number of non-linear algorithms are described thoroughly in (Bazaraa et al., 1993; Train,
4.6 Simulation based techniques

The choice of the optimisation procedure should reflect the complexity and behaviour (global concave, locally non-concave, point-wise not twice differentiable) of the likelihood function to be maximised. When linear and non-linear algorithms have to give up, the development within computers have made the way for simulation based methods.

4.6 Simulation based techniques

There are at least two reasons for applying simulation when estimating a model. First, the model may include stochastic elements for which parameters only can be retrieved this way and secondly, the function may be of a such complex character (non-closed form) leaving out the possibility of direct estimation. The general approach applied in recent modelling MSL, is described in the following section whereas more general simulation techniques follows.

Two important characteristics of simulation techniques are unbiasedness or consistency and efficiency. When the estimator $\tau$ of $t$ bring out the ‘true’ value the estimator is said to be unbiased (technically, $E(\tau) = t$). When the variance of the estimator is the least possible the estimator is said to be efficient (no estimator $\tau'$ for which $\text{var}(\tau') < \text{var}(\tau)$, exists).

4.6.1 Maximum simulated likelihood

Model types allowing for non-deterministic behaviour by inclusion of additional stochastic elements, cannot be estimated by the methods described in section 4.5. Given a specification of the utility that includes distribution of stochastic terms, the likelihood function can formulated and solved by Maximum Simulated Likelihood (MSL). However, MSL does involve one integral $(-\infty, \infty)$ for each dimension of the distribution of the added stochastic term(s), plus the one inherited from the residual term.

Following the specification of the likelihood function in section 2.2.1, by which

$$L = L(\beta|X) = \sum_n \sum_i g_{in} \ln(P_i). \tag{2.3}$$

where $\beta$ are coefficients and $g_{in}$ is 1 if alternative $i$ is chosen, 0 otherwise; $P_i$ is the probability for choice of alternative $i$ under some probability model e.g. MNL.

Adding further distributed terms to the utility, corresponds to letting $\beta$ be distributed $f(\beta|\theta)$ where $\theta$ is the parametrisation of the distribution. Let $\beta^r$ refer to the $r$th of $R$ draws from $f(\beta|\theta)$ and $P_i(\beta^r)$ the probability for choice of alternative $i$ given the drawn value $\beta^r$. Averaging over successive draws from
CHAPTER 4. DATA AND ESTIMATION

\[ f(\beta | \theta), \text{ gives} \]
\[ \hat{P}_i = \frac{1}{R} \sum_n \sum_j g_{jn} \ln P_j(\beta^r). \tag{4.7} \]

The simulated log-likelihood is then given by

\[ SL = SL(\theta | X) = \sum_n \sum_j g_{jn} \ln \hat{P}_j, \tag{4.8} \]

and the estimator for the coefficients is the \( \theta^* \), for which \( SL \) obtains the maximum value. As the maximum likelihood estimation is conditional on the formulation of the probability function \( P_i \), the simulated maximum likelihood estimation is conditional on the formulation of the probability function \( P_i \) and the functional form assumed for the distribution.

Hajivassiliou and Ruud (1994) derive properties of the asymptotic distribution, e.g. that the MSL estimator is consistent if the number of draws for each observation rises with the number of observations.

One criticism of the MSL is that it only optimises parameter values of the specified a priori distribution. In section 7.1, it is shown, that even for erroneous specified distributions, (significant) parameter estimates are obtained along with (significant) estimates of the ordinary utility coefficients.

### 4.6.2 Quadrature

Probit models (and other models involving only few distributed terms) may be solved by means of quadrature. The most frequently applied are Geweke-Hajivassiliou-Keane (GHK) and Gauss-Laguerre (GL) quadrature. Differences between the quadratures are determined by the weight functions. The methods has been left behind for the more efficient MSL method involving pseudo-random or Halton draws, described in section 4.6.3.

### Method of Simulated Moments

This method (MSM) is applied to simulated the probabilities of e.g. the multinomial probit model (MNP), which cannot be estimated analytically due to the integrals. In section 3.5.9, the MNP model was specified by the utility of the alternatives. The following specification is based on the utility differences, i.e. \( \tilde{U}_j = U_j - U_i, i \neq j \). Elements \( \tilde{u} \) of \( U \) are normal distributed since the normal distribution is additive\(^3\)

\[ P(i) = \int_{-\infty}^{v_1} \int_{-\infty}^{v_2} \ldots \int_{-\infty}^{v_n} f(\tilde{u} | \Omega) d\tilde{u}_n d\tilde{u}_{n-1} \ldots d\tilde{u}_1. \tag{4.9} \]

This can be further rewritten by introduction of a density function \( h(\tilde{u}) \);

\(^3\)For \( X, Y \) distributed normal, \( aX + bY \) is distributed normal for \( a, b \) real.
4.6 Simulation based techniques

\[ P(i) = \int_{-\infty}^{v_1} \int_{-\infty}^{v_2} \cdots \int_{-\infty}^{v_n} f(\tilde{u}|\Omega)h(\tilde{u})d\tilde{u}_n d\tilde{u}_{n-1} \cdots d\tilde{u}_1 \]  
\[ = E\left( \frac{f(\tilde{u}|\Omega)}{h(\tilde{u})} \right). \]  

(4.10)

MSM is consistent for a fixed number of repetitions when the weights are independent of the residuals. In practice, this requires drawing from two densities.

The Method of Smoothed Simulated Moments

The Method of Smoothed Simulated Moments (MSSM) is also known as Stern’s Method (Stern, 1992). The method is suggested as an improvement of the MSM, as this reduces the variance on the simulated probabilities but maintains the unbiasedness.

Assuming the error term is normal distributed, whereby the utility difference is normal distributed, the choice of an alternative \( i \) from a set is described by

\[ P(i) = P(U_i \geq U_j, \forall j) \]
\[ = P(\tilde{U}_j \leq v), \]

(4.11)

where \( \tilde{U}_j = U_j - U_i \) and \( v \) is the difference in the stochastic parts of the utility. Further the utility is partitioned into \( \tilde{U} = Y + Z, Y \sim N(0, \Lambda) \) and \( Z \sim N(0, \Omega - \Lambda) \).

\[ P(i) = P(\tilde{U} \leq v) \]
\[ = P(Y \leq v - Z) \]
\[ = \int P(Y \leq v - z|z)f(z|\Omega - \Lambda)dz \]
\[ = \int \left\{ \prod_j P(Y_j \leq v_j - z_j|z_j) \right\} f(z|\Omega - \Lambda)dz \]
\[ = \int \left\{ \prod_j \Phi(\frac{v_j - z_j}{\sqrt{\Lambda_j}}) \right\} f(z|\Omega - \Lambda)dz. \]  

(4.12)

Equation (4.12) is the MSSM estimator \( (P^*(i)) \), which is calculated by draws from a normal distribution and evaluation of the product. The method can be improved by either better choices of \( \Lambda \) or a larger number of draws from the normal distribution. The method can be generalised to the GHK estimator.

Alternatives are the Method of Simulated Scores (MSS), the Accept-Reject (A-R) estimator, the Smoothed Accept-Reject Simulator or the Simulator Crude Frequency Simulator. Markov Chain Monte Carlo Methods, in particular Gibbs sampling or simulated annealing, can perform estimation of the general likelihood function. Simulation assisted methods are further described in Train (2003).

\(^4\)The covariance matrix \( \Lambda \) is diagonal and \( \Omega - \Lambda \) positive semi-definite.
4.6.3 Random numbers
An underlying concept of simulation methods is random number generation. These may be generated by various means, where probably the most common is pseudo-random numbers. Numbers are generated to reflect the distribution, preferably without correlation between successive draws. The drawback of the method is the often quite large number of draws, hence estimation time, required (for a description of techniques see e.g. Ross (1997)).

An alternative to random numbers, may be Halton sequences, where (negative) correlation is introduced. According to Bhat (1997), this significantly reduces the number of draws required to obtain the same accuracy in the estimation. At this point it should be mentioned that the comparison in Bhat (1997), is between Halton sequences and simple random number generators, which have been proven to produce serial correlation. Had the comparison been with ‘second generation’ random number generators, which produces one random number by use of draws from two different distributions (such generators are free of serial correlation and have a much longer period), the issue could have been settled. These ‘second generation’ generators are further described in Press et al. (1994), while the impact of a poor random number generator on traffic estimation is discussed in Sørensen et al. (2001).

4.7 Sample based techniques
The above description of estimation and estimators applies conditional on existence of the estimator, that is, we have an expression for calculation of the estimator. Unfortunately, this is not always the case; in these remaining cases simulation can be applied. Two examples of techniques that can be applied are the Jackknife and the Bootstrapping methods, which are described below. The purpose of the description here is to enable a comparison later in the thesis and not for actual application.

4.7.1 Jackknife
One of the earliest techniques to obtain reliable statistical estimators is the jackknife technique. It requires less computational power than more recent techniques (e.g. Bootstrapping). For the jackknife, the data set is divided into subsets and the statistic of interest is computed for each of the subsets. The average of these subset statistics is compared with the corresponding statistic computed from the entire sample in order to estimate e.g. the variance of the mean.

Formally, let $X$ denote a sample of $n$ units, let $\hat{\theta} = s(x)$ be the estimator of interest. For the Jackknife estimation re-samples are drawn from the sample, where exactly one observation is left out compared to the original data set; i.e. $X(i) = \{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N\}$ for $i = 1, \ldots, N$. 
4.7 Sample based techniques

For each of the $N$ new data Jackknife samples (re-samples) an estimate of $\hat{\theta}$ can be calculated; for the $i$th data sample this is denoted $\hat{\theta}_{(i)} = s(x_{(i)})$. The Jackknife estimate of $\hat{\theta}$ is the average of the estimates for each sample $\hat{\theta}_{(i)} = \frac{\sum \hat{\theta}_{(i)}}{N}$.

The Jackknife only works well for linear statistics (e.g., mean). It fails to give accurate estimation for non-smooth (e.g., median) and nonlinear (e.g., correlation coefficient) cases. The most common use of the Jackknife is to estimate the bias of a statistic.

Delete-a-Group Jackknife

A variant of the Jackknife is the Delete-a-Group Jackknife (DGJ) where several data observations are deleted at a time, rather than one observation. The principle is that the data sample $X$ is systematically divided into a relatively large number $G$ groups of equal size and the Jackknife samples are then formed by deleting one group of data from the sample at the time, whereby $G$ such re-samples are formed. The DGJ-estimate is obtained as the weighed average over estimates from each sample; the weight reflect potential differences in re-sample sizes. The procedure is further described in Kott (1998).

4.7.2 Bootstrapping

The idea behind the method of Bootstrapping is to enable the estimation of some statistical properties, which is not immediately feasible given the available data.

Given a data set $X$ of $N$ units, $B$ samples of $N$ units are drawn at random, with replacement, from the data set $X$. The number of 'new' samples (re-samples) $B$, should be a fairly large number e.g. 1000, 2000, 10,000. It requires that the sample is a good approximation of the population. The method does not work properly for small sample sizes ($N$ should be more than 20). An example of how the re-samples are formed is shown in table 4.1 (the data set does only consist of 7 units for the sake of illustration). Since the re-samples are drawn with re-

<table>
<thead>
<tr>
<th>Original data $X$</th>
<th>1 2 3 3 5 6 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resamples $b = 1$</td>
<td>1 2 5 2 2 5 6</td>
</tr>
<tr>
<td>$b = 2$</td>
<td>2 5 2 2 5 5 6</td>
</tr>
<tr>
<td>$b = 3$</td>
<td>1 5 1 6 3 2 5</td>
</tr>
<tr>
<td>$\vdots$</td>
<td></td>
</tr>
<tr>
<td>$b = B$</td>
<td>5 1 3 2 3 5 6</td>
</tr>
</tbody>
</table>

Table 4.1: Example of re-sampled data sets used in Bootstrapping.
placement, the number 2 can appear in one of the new samples more times than it appeared in the original sample – this can be seen in $b = 2$. Similarly, not all numbers in the original data set is expected to be found in the new samples ($b = 1$).

Hereby, we get $B$ re-samples rather than one sample of size $N$, from which we can obtain the estimates, as if they were samples drawn directly from the population. Stated differently, the sample (the original data sample) is treated as if it was the population (and not a sample drawn from it). For each of these re-samples, the statistic of interest is computed for each re-sample and the empirical distribution is formed (plot the estimated values in a histogram to find the distribution). From this ‘Bootstrap’ distribution, statistical inferences can be performed, e.g. hypothesis testing.

The Bootstrap method can be used to find the Bootstrap distribution for any well-defined function of the sample data, without making any assumption of the population the sample it is drawn from, other than that the sample itself must be drawn at random. Further, as the sample size $N$ goes to infinity, the Bootstrap sampling distribution approaches the ‘true’ sampling distribution for the population.

Other variants of the Bootstrapping method do exist and are frequently used to analyse confidence bounds for networks and neural networks. The bootstrapping method is further treated in Efron and Tibshirani (1993), though it was first presented in Efron (1979).

4.8 Forecast

The purpose of estimating a model may be twofold; to obtain information on users behaviour and to enable a forecast calculation. This section is concerned with the latter of these issues.

Common for the two issues is the aim at estimating the correct (or ‘true’) coefficients, if possible. Estimation of the correct coefficients requires that the correct model is specified. But in the real world application, we do not know whether we use the correct model or not (we hope that we do) and appreciate that the residual term seek to capture the remainder of the variation in the data.

To illustrate the importance of estimating the correct coefficients an example has been constructed. The choice between three alternatives (A,B,C) is described by three variables $X_1, X_2, X_3$ each multiplied by a coefficient $(\beta_1, \beta_2, \beta_3)$. In the case of forecasting, the coefficients are obtained from some data set and are used to predict the probabilities for choice between alternatives, given a description ($X$ values) of the alternatives. Three different $\beta$ vectors are constructed, such that the first is the ‘true’ coefficients. The second, to resembles the results of a
4.8 Forecast

<table>
<thead>
<tr>
<th>Alternative</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$P(i)$ for different $\beta$ vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>20</td>
<td>11</td>
<td>$B^1$ = $(-1, -1.5, 1)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$B^2$ = $(-1, -1.6, 1.2)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$B^3$ = $(-2, -3, 2)$</td>
</tr>
<tr>
<td>B</td>
<td>12</td>
<td>19</td>
<td>11</td>
<td>0.55</td>
</tr>
<tr>
<td>C</td>
<td>20</td>
<td>25</td>
<td>27</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Table 4.2: Differences in probability for choice of each alternative, caused by small changes in coefficients.

Let the first vector of coefficients $B^1 = (-1, -1.5, 1)$ be the ‘true’ coefficient and thereby the basis for comparisons. The second vector $B^2 = (-1, -1.6, 1.2)$ has a numerical increase in second and third coefficient, which causes change in all ratios of coefficients. The third vector $B^3 = (-2, -3, 2)$ is the double of the first – which suggest that the coefficients were estimated from a poor data set, in the sense that the correct trade-off between coefficients was estimated but with a high scale – recall, that $\mu\beta$ and not $\beta$ is estimated.

The probability for choice is evaluated by the MNL (described in section 3.3) for three different $\beta$ vectors and results are given in table 4.2. According to the probability for choice of alternatives evaluated based on the correct coefficient values $B^1$, slightly more than half would chose A (55%), a third would chose B (33%) and one in every eight would chose C (12%).

A small change in the coefficients (from $B^1$ to $B^2$) actually caused a major change in the choice probabilities. The most likely alternative (A) given the correct coefficients ($B^1$), is now only expected to be chosen by one in every five (20%)! Further, the least likely alternative given the correct coefficients (C), has become the most likely and is expected to be chosen by two in every three (67%)! Choice of alternative B, has changed from a likely choice of one in three, to one in eight (13%).

Doubling the coefficients also had an effect, though of smaller scale than the change of relative coefficients. The order of ‘most likely choice’ is not affected. However, the probabilities did change. The most likely choice given the correct coefficients (A), became even more likely (increase from 55% to 71%). The remaining two alternatives B and C, both experienced a decreasing likelihood for choice; B decreased from 33% to 26% and C decreased from 12% to just 4%.

The change in probabilities for choice of alternatives when the coefficients are
all doubled, is explained by the exponentiation of the utilities performed in the calculation. Since the exponential function is a positive, monotonic function, the order of the values is unaffected by it. The change in probabilities is due to the non-linearity of the exponentiation.

As this example clearly shows, it is not without importance to estimate the correct trade-offs between variables (ratios of coefficients) but also to estimate the correct level for the coefficients (including correct logit scale). If e.g. a new mode is introduced and the expected market share is calculated this way and the design e.g. number of lanes on a new road or the headway, is determined based on this, we have a problem!

### 4.8.1 Segmented model

The example just described did not include segmentation of data, though this is often the case in applications. Adding data segmentation is a relatively simple task, and is described below.

In a segmented model (described in section 4.3) the probabilities for choice of an alternative $i$ based on the data segmentation $X^1, \ldots, X^M$, is given by equation (4.3), reproduced here

$$P(i|\beta^1, \ldots, \beta^M; X^1, \ldots, X^M) = \sum_m P(m) P(i|m; X^m). \quad (4.3)$$

The choice probability for each alternative is weighted by the probability for membership of each segment $P(m)$.

As the model involves two terms $P(i|m)$ and $P(m)$, which may evolve over time, both these have to be forecasted. The importance of correct estimates of the utility coefficients has been emphasised above. The importance of correctly predicting the membership probabilities is a matter of weighting the probabilities choice of each alternative $P(i|m)$, correctly. Hence, the importance of forecasting the membership function correctly is evident.

For forecasting in general, a change in the estimated coefficients or the formulation of the utility coefficients are seldom expected, that is, the $P(i|m; X^m)$ is expected stable over time. However, one example of coefficients changing over time has been found, where 'barrier of integration' and an 'economic barrier' towards travel over the recently opened bridge between Denmark and Sweden were included. These barriers were incorporated as an out phasing negative effect, see Øresundskonsortiet (1999).
4.9 Summary

The purpose of this chapter was to indicate how important the process before and after the model selection is for the final result. When a study of the impact of some change in the supply of transport is about to be constructed, several steps are important. The area of interest, the alternatives of interest, data availability, data validation, choice of discrete choice model, model estimation, model validation and finally presentation of results. This thesis has covered some of the issues; namely the data collection, storing/handling data, segmentation and/or merge of data; range of probability models applicable for discrete choice and choice of model (in chapter 3) and finally, estimation techniques including simulation assisted techniques.

Estimation of a model can be partitioned into three steps; obtaining data including data validation; setting up the formulation to be estimated and the actual estimation process. The model formulation was the subject of the previous chapter – this chapter addressed the data collection (whether it is by questionnaire or from previous studies), data segmentation and the estimation part.

As an initial step the terminology for describing data was established, ranging from population to the smallest unit. This was followed by a discussion on data quality. Data of 'high quality' is described as a large data set where question and answer match, where the accuracy in the answers is high, numerous (relevant) socioeconomic attributes of travel, a variation of choice of alternative, etc. Lacking just one of these characteristics may severely influence the quality of the data. Different sources of bias and the impact on the modelling task were discussed.

Data collection and handling is a separate issue closely related to forming a model, but also in itself growing to be an independent research area. The most common types of data collection methods; Revealed Preference (RP), Stated Preference (SP) techniques, Transfer Price and Rank Order data were described.

Probably the most important lesson to be learned about data, is to be very carefully when collecting, handling and employing it! Bias may arise virtually anywhere in the process and only some types of bias can be identified and accounted for at later stages in the modelling. Saying this, is also saying that any bias caused by collecting or handling of data is inherited throughout the model, as no matter how ‘perfect’ a model is, it cannot patch up for bias in data.

\footnote{At several international conferences, separate sessions are arranged for data collection methods/stated preference studies. The development within this area seems to have speeded up recently, wherefore recommendations in e.g. Ben-Akiva and Lerman (1985); Richardson et al. (1995) may have been overtaken. As data collection methods are not the core issue of this thesis, further discussions of this issue have been refrained from.}
CHAPTER 4. DATA AND ESTIMATION

A trick often applied when setting up a model is to segment data. Prior to discussing the various criteria for segmenting data, the terminology was established, followed by a discussion of existence. The list of possible criteria for data segmentation can be very long, in this thesis some examples have been given. The section also gives a description of different techniques for selecting/drawing a sub-sample from a data sample. The methods covered rigid methods for selecting a sub-sample e.g. strata proportional to a (few) variable values. Samples can also be drawn by use of randomised techniques; where members are drawn randomly with/without replacement (random samples), which may be combined with stratification, whereby a stratified random sample is obtained. Further, cluster analysis can be used to select sub-samples conditional on values of variable or functions hereof. Clusters are selected such that close observations (units) are grouped together, distant observations are assigned to different clusters. Different distance measures were described.

The estimation techniques considered in this chapter, are currently applied within the area of transport demand modelling. As models grow complex, complex measures must occasionally be considered when models are estimated. The simple MNL model is on closed form and can therefore be estimated by the Newton Raphson (NR) algorithm. Though, for MNL models based on many attributes other algorithms may converge faster. When the residual terms are no longer independent, the models are not on closed form, wherefore simulation based techniques must be employed. Again, depending on the actual structure, different simulation techniques may converge faster. Models not involving other distributed terms than the residual can generally be handled be the BHHH algorithm. If more distributed terms are included e.g. as in the Mixed Logit, the maximum simulated likelihood method (MSL) can be used. A great deal of research is being done on this subject, wherefore no clear cut conclusions were given on which method is the most efficient for each model. Rather, an example of a method that works in general, was given.

A prerequisite for closed form estimation (NR, BHHH) is the assumption of existence of estimators. Further, tools to apply are the simulation based techniques (MSL, quadrature) and the sample based techniques. Two different techniques (Jackknife, Bootstrapping) were described in the thesis – not because the models described require such solution techniques, merely for comparison with techniques proposed later in the thesis. The two techniques require a random sample from the population, which is subsequently treated as if it were the population. Multiple new samples (re-samples) are drawn from the sample and estimates can be obtained by comparisons of estimates obtained for each of the samples.

The motivation for undertaking data collection, model selection and estimation procedure is often forecasting. This chapter concluded with describing how to perform a forecast and a brief discussion on the importance of performing forecasts based on the correct model and correct (absolute) coefficient estimates,
4.9 Summary

for both the un-segmented and the segmented model.
Chapter 5

Data segmentation by preferences (SEP Method)

5.1 Introduction

During the past two decades a significant amount of effort has been put into improving the models with a special focus on the (theoretical) model structure. Recently, a great deal of the research has been directed towards error component models where differences between individuals have been modelled by means of statistical distributions. Differences between individuals have been realised some time ago, and has generally been incorporated by means of data segmentation. Basically, segmentation is a division of data into sub-samples, where separate models are estimated for each segment. Observations may be segmented by e.g. the purpose of the trip, income levels or any other readily accessible variable.

Different models (coefficient estimates) for each segment enables the models to capture differences between segments (groups of individuals), that is, enables individuals in different segments to be described by different relative preferences. However, the individuals within each segment are still assumed identical in preferences (and IID error term depending on model structure, see section 3). In other words, it is recognised that some individuals are different (the urge for segmenting the model) but at the same time, that individuals within the same segment, are assumed identical in preferences. This is why care should be taken when segmenting data!

Differences between individuals may be modelled by the random coefficients logit model (RCL), which introduces a stochastic component into the coefficients (preferences) the individual is described by, see section 3.6.1). The case for estimation of models, is that RCL may be regarded as an alternative to segmentation or may be used on top of the segmentation, with differently dis-
distributed stochastic terms for each segment – or even distributions added to different coefficients. Differences in distributions may appear as different coefficients, or even different shape of distribution.¹

If the model considered is to be used for *forecasting*, there is a somewhat clearer difference between using segmentation and a RCL specification. In a segmented model, variables related to the segments must be extrapolated to the time of the forecast. These variables include socioeconomic variables, attributes of transport alternatives, size and composition of the segments (e.g. some segment of individuals may change in relative size). For the RCL model, where the distribution of the coefficients is not directly linked to the individuals, changes in the overall distribution is at best, quite difficult to calculate. Changes may include increased/decreased variance and/or mean, and even change in shape of distribution.

During the formulation of a model it is implicitly assumed that all individuals act *independently* of one another (all IID models) and may be described by the same utility function (all models). Some alterations of the latter have been put up, by introducing heteroscedasticity (see e.g. Bhat, 1995 or Bhat, 1997) or by introducing stochastic preferences (in addition to the stochastic error term \( \epsilon \)) into the models. This, of course, is put on top of the traditional segmentation by purpose, which implies different models for each purpose; possibly with restrictions on coefficients in the various models. To the author’s knowledge, little has been done to analyse the effect of the segmentation, that is, how to produce the segments as well as the interaction between segmentations on one hand and models with stochastic preferences (heteroscedasticity and error components) on the other hand. This section proposes a method by which segmentation of data can be obtained from the data – especially from the individuals’ own preferences. Preferably by use of a segmentation criterion also applicable from a planning or strategic point of view.

### 5.1.1 Background

The vector of coefficients \( \beta \), is estimated by maximising the log likelihood function \( \mathcal{L} \). For description of the segmentation method the MNL model is applied, leading to the following log likelihood function to be maximised.

\[
\mathcal{L} = \sum_{n} \sum_{i} g_{in} \ln(P_i), \quad \text{where} \quad P_i = \frac{\exp(V_i)}{\sum_{j} \exp(V_j)},
\]

¹Applying a RCL formulation to a segmented model, may be replaced by a RCL on an un-segmented model, where the distribution captures the density mass of the latent segments. The shape of the distribution may be multi-modal with each mode corresponding to a segment, and likely being local non-symmetrical. Hence, adding RCL on top of a segmented model seems to be a more applicable method, though inter-relations between preferences and segments are left unaccounted for.
5.1 Introduction

where \( g_{in} \) is 1 if alternative \( i \) is chosen by individual \( n \), zero otherwise. \( P_i \) is the probability for choice of alternative \( i \) with deterministic utility \( V_i \). The log likelihood function for a segmented model with \( n_t \) individuals in segment \( t = 1, 2, \ldots, T \), is given by

\[
\mathcal{L} = \sum_t \sum_{n_t} \sum_i g_{in} \ln(P_i),
\]

that is, the overall likelihood is the sum over likelihood for each segment. Hence, overall accuracy of the model (in terms of likelihood value), can be improved if the accuracy (likelihood) for each of the segments’ model is improved.

A searching algorithm that optimises the likelihood function performs the actual optimisation, that is, maximises the probability that the observed values could be realised (given the parameterisation and the data). At this point it should be stressed, that in principle, each individual has it’s own true vector of coefficients, whereas the estimation procedure seeks to assimilate these within data samples.

Segmentation methods can be roughly divided into two groups; those based solely on the (raw) data and those based solely on the coefficient estimates \( \hat{\beta} \). The first group contains, among others, the traditional segmentation by purpose which distinguishes itself by low additional estimation time\(^2\) compared to the unsegmented model. At the same time the quality of the segments seems to be reasonable. To this, adds the rather obvious, advantage of ease of interpreting the segments. The second group of segmentations, is based on analysis of homogeneity in the coefficients. The segmentation is constructed such that the similarity between individuals within the segments is maximised.

To improve the model, the explanation by the deterministic part of the utility must be increased whereby less variance is described by the stochastic part (the residual). This requires that the coefficients are estimated as correctly as possible, i.e. the variation on the coefficient estimates is minimised and the specification must be as close to the 'true' specification as possible\(^3\). That, in turn, leads to that segmentation has to be constructed such as to minimise the range of the (true) coefficients \( \hat{\beta} \) within each segment. Hence, the segmentation must be constructed such that individuals with highly different (true) coefficients are allocated to different segments while the individuals with the most similar (true) coefficients are allocated to the same segment. However, application of this to a given data set would normally not be possible due to the significant amount of data required to estimate the vector of coefficients for each individual.

\(^2\) Time used for the estimation is less than proportional to the number of segments.

\(^3\) Though we don’t know the 'true' specification of the utility function, this search is limited by available data for the specific context.
5.2 SEP method for data segmentation

In this section a method for SEgmentation by Preferences (SEP) is proposed. Further, the SEP method is motivated and existence of segments formed by the SEP method is discussed.

5.2.1 Purpose of the SEP method

The purpose of the proposed method is to detect a data segmentation where the within-segment variation is minimal. The method seeks to determine such a segmentation based on the revealed preferences obtained from model estimations within a discrete choice framework. Preferably, we would like to estimate a model for each data record to facilitates the minimal variance segmentation. However, estimations must be performed on smaller subsets of data, since model estimation is not possible to perform for each individual. A model estimation requires more data records than the number of coefficients to be estimated; this is a matter of identifiability of the models, which was described in section 2.3.2. Generally, a data set with multiple records per individual is not available wherefore the choice here has been to estimate models based on sub-samples of the data (or groups of records).

In the proposed method, data is analysed such that (groups of) individuals highly alike in terms of preferences, will be grouped together and at the same time, (groups of) individuals with highly different preferences are allocated to different segments.

A one-coefficients example of this is shown in figure 5.1 where the red line is the mass density of individuals coefficients, for the ‘true’ coefficient values on the horizontal axis. As the figure indicates there is a large faction of the individuals who’s ‘true’ β is close to a ‘low’ value (βlow), and an other large faction of the individuals who’s ‘true’ value is close to a ‘high’ value (βhigh). Segments are formed according to peaks in the density mass, as shown below the horizontal axis.

The mass of the first peak is larger than the second, indicating different number of individuals belonging to each segment (segment A is larger than segment B). The principle is similar for higher dimensions, where segments are identified as areas with a high density mass of the simultaneous distribution for the vector of coefficients.

5.2.2 Motivation for the SEP method

For the sake of illustrating the motivation for the SEP method, assume that the ‘true’ values for the coefficients in the utility function are known.

In figure 5.2, the mass density of the ‘true’ coefficients is drawn by a contour
plot. Two different segments are illustrated where the shape of the distribution may differ and the density for the segments may overlap, as illustrated. Implicitly, it is assumed that the 'true' coefficients are distributed – otherwise the mass density would be degenerate at the common 'true' value. The shape of the mass distribution is an example – this was chosen for the illustration of possible differences between their distribution, and not as a specific guess/intuition of the shape of the distribution. The coefficient estimates for each segment are marked by $\times$; coefficient estimates are not necessarily at the centre of the density distribution, as shown in the figure.

That coefficient estimates are different for different segments, is not the same as assuring that the mass density of 'true' coefficients does not overlap. Consequently, two records with the same 'true' coefficients are not assured to be members of the same segment; this is evident from this example.

The purpose of segmentation is to get at better description of the behaviour of a collection of individuals. This is done by decreasing the within-segment variance on the coefficient estimates, or more generally speaking, individuals (data records) with similar behaviour are assigned to the same segment. Looking at figure 5.2, it is obvious that a better partitioning of data will lead to a smaller variance of the 'true' coefficients, for each of the segments. This can be done by switching segment membership, such that the density distributions do
not overlap.

Simplifying this idea by not having to draw the density distribution for each segment, segments can be formed by segmenting solely by the 'true' coefficients. This is illustrated for two coefficients ($\beta_1, \beta_2$), in figure 5.3. Two points on a line through the Origin have the same ratio of coefficients (relative values). The difference between two points on the same line through the Origin is that the logit scale increases the further away from the Origin the point is.\footnote{Note, that coefficient estimates scaled by the logit scale factor $\mu\beta$ are estimated rather than the coefficients $\beta$ themselves.}

Dotted lines through the Origin, separates the segments such that two different data records with the same coefficients are ensured to be in the same segment. Further, two different data records with the same coefficient ratio are ensured to be in the same segment, as these would be on the same line through the Origin (the ratio is the slope of the line). In higher dimensions (multiple coefficients), segments are shown by cones with the acute end in the Origin.

Common for (traditional) segmentation by e.g. purpose and segmentation by preferences is the unambiguous division of data into segments. The difference, however, is how this is performed, and thereby which records are members of which segment. Consequently, the description of the segments are different and the coefficient estimates are likely to be different.
5.2 SEP method for data segmentation

![Coefficients space for segments A, B, and C with estimated coefficients](image)

**Figure 5.3:** Three segments defined by similar coefficients.

### 5.2.3 Existence

Having defined segments as a division of a sample of data into a set of totally exhaustive and mutually exclusive sub-samples, the issue of existence of segments follows naturally. By existence, is meant whether it is actually possible to find such a division of data, where individuals allocated to each segment behaves homogeneously, and where potential changes in individual behaviour over time is consistent within each segment (and hopefully may be described). Preferably, the segmentation should be both logical and meaningful.

The criterion of homogeneity of preferences within segments, arises from the choice model, where it is assumed that individuals can be described by the same utility function, the residual terms captures the deviances.

The second criterion of consistency of change of preferences within each segments, is required if the model is intended for forecasts and that changes over time can be described (if not, how can a forecast model be based on them?). Such verifications is more complicated, and may be performed by use of time series data. By describing segments for all time periods, and preferences for each of the time periods, (possible) changes in preferences over time, may be captured\(^5\). Hereby, enabling the modeller to forecast the changes in preferences. In the following, existence of segments is assumed.

---

\(^5\)The task of describing changes becomes more complicated if the preferences (for each time period) are not described in a fairly simple manner.
5.3 Algorithm

The SEP method can be implemented by use of the following algorithm. The algorithm requires a data set with covariates for each of the alternatives and data on which alternative is chosen. Further, it is assumed that individuals within the same segment are described by the same model (choice described by the same utility function) and by the same vector of coefficients. The traditional assumption is that the same model describes all individuals.

The algorithm shown below and subsequently, each step is explained in further detail.

**Step 1** Initialisation: Set number of groups $K$ and grouping variables. Estimate the unsegmented model (inclusive likelihood $\mathcal{L}$).

**Step 2** Generation of $K$ data groups based on the grouping variables.

**Step 3** Estimate the vector of coefficient $\hat{\beta}^k$ for each of the $K$ groups.

**Step 4** Find clusters based solely on the coefficient estimates $\hat{\beta}^k$ from step 3.

**Step 5** Identification of segments from clusters. Estimate likelihood and test statistics for the segmented model.

The initialisation (step 1) serves to set the stage for the determination of the segments. One of the keys to the method is the division of the data sample into groups; the number of groups is related to the number of observations in each group. The model estimation requires some minimum number of observations before the estimators for the coefficients are identified (described in section 2.3.2). Though, at the same time the number of groups should be as large as possible to enable the model estimation to detect possible differences between observations (larger groups tend to pull the coefficient estimates towards the estimates for the model based on the whole sample).

After initialisation step 2 proceeds with an initial grouping (data partitioning) of the data. This partitioning can be by cluster analysis or simple division proportional to a variable. At this point it should be noted that the initial groups should be neither too small nor to large. The first case will lead to inability to estimate models for the groups while the latter will make the assumption of homogeneity within the groups become more critical. In general, the number of groups should be maximised with regards to that no group size is below some threshold. If the initial groups are generated by cluster analysis, the group size is subject to choice of generating algorithm and may vary between
5.3 Algorithm

groups. Group generation is described in more detail in section 5.3.1.

Model estimation for each group follows in step 3. Models are estimated using the same type of model (e.g. MNL) and specification of utility function. If the $t$-statistics for some group, indicate that a coefficient should be excluded, this is ignored, as this will lead to a larger error, than the lack of significance of one coefficient at this stage of the model will give rise to.

Based on the coefficient estimates, clusters (of groups) are constructed by cluster analysis (based solely on the coefficient estimates), in step 4. The objective is to identify where the density mass is for the simultaneous distribution of the coefficient estimates. All groups with similar estimates are grouped together; different clusters have different locations (in the space defined by the coefficient estimates). In particular, if two different groups have almost identical coefficient estimates, they are not assigned to different clusters. The number of clusters can either be predefined (maximal number) or be identified along with the clusters.

The purpose of the final step 5 is to identify segments, that are largely indistinguishable from the clusters. Segments are to be described by a (few) variables in the data. However, this might involve that some observations belongs to same cluster but different segment. This is exactly how the distinction between clusters and segments arises. To get a grasp of the cost (in terms of model fit) of this translation, a model estimation can be performed based on the clusters. The likelihood and $t$-statistics for the cluster based and segment based models, are then compared to see the effect of the translation from clusters to segments, on the model fit. This step is further discussed in section 5.3.2.

The segments are applicable in other contexts, eg. in a forecast model or for strategic planning/ development of a market if price differentiation is in question. Direct marketing may become more attracting as the reaction, for each segments, to a given change in the attributes (change of price, information available) is known.

By application of the algorithm, the total number of estimated coefficients in the model is increased by a factor $K$ (number of segments). This has a negative impact on the test statistics. Hence, if this method is applied to a smaller data sets the number of resulting cluster $K$ should be small or restrictions between coefficients in different segments should be introduced, ie. only some coefficients vary between segments.

---

6The statistics package SAS© was used for the data analysis and cluster formation. No option for 'minimum size of clusters' was found, wherefore cluster methods that tend to produce fairly equal sized clusters were preferred. It is beyond the scope of this work to program a cluster method where such a criterion is included.

7This can be automatised by setting a level for distance between clusters, in the cluster formation procedure.
CHAPTER 5. DATA SEGMENTATION BY PREFERENCES (SEP METHOD)

5.3.1 Data grouping

A key issue of the SEP method is the grouping of data. Some systematic approaches suggest themselves such as the method of random sampling, partitioning proportional to some variable as well as cluster analysis.

By grouping the data sample \((X)\) a division of data into \(K\) smaller groups \((X^1, X^2, \ldots, X^K)\) which are mutually exclusive \((X^k \cap X^j = \emptyset, j \neq k)\) and totally exhaustive \((X = \bigcup_k X^k)\) is referred to. Groups are allowed to be different in size, though non-empty \((X^k \neq \emptyset, \forall k)\).

Using random sampling to generate groups has the advantage of preserving the statistical properties of the data e.g. the mean value of the sample. Further, it is very easy to fix the size of the groups and the total number of groups. Observations within each group may be very similar (by chance!), though they are most likely to be different, in terms of values of descriptive values. Random sampling was further described in section 4.4.1.

Cluster analysis is a tool to partition a data sample into sub-samples by a measure of similarity, such that similar observations are grouped together and very different observations are placed in different groups. Groups formed in this manner are referred to as 'clusters'. The maximal number of clusters is easily fixed, though the maximal size of each cluster is difficult to control.\(^8\) Cluster analysis was further described in section 4.4.2.

At this point it should be kept in mind, that the purpose of the method is to find patterns of similar preferences across the individuals in the sample. Further, recall the formulation of segmentation in section 4.3: \textit{A data segment is a subset of a sample, which is unambiguously described on the basis of a (few) simple criterion(s).} The segments we are aiming at should be possible to describe by relatively simple rules (above/ below some threshold value), wherefore very 'diverse' observations should not be assigned to the same group. The diversity between observations is to be seen as differences in the descriptive variables (including the variables used in the utility function and any other available variables).

For these reasons application of cluster analysis to form the groups is suggested. Deciding to use cluster analysis is no the last decision to make. As referred to above, a measure of difference between observations is to be used. This requires two further decisions; which measure of distance to employ and how many and which variables should the measure be based on?

Selecting one before a number of measures of distances (the list of measures is shown in section 4.3) is not without complications, as none of the measures

\(^8\)For this study cluster analysis tools in the statistical package SAS were used. Neither of the available procedures in SAS, had an option to restrict the maximal or minimal size of clusters.
5.3 Algorithm

have been proven superior to the other, in any aspect! If there is an expectation of which shape the (circle, non-convex, elongated, etc.) groups are expected to have, some of the measures are known to perform better than the others. In this case, no such prior knowledge is available wherefore the recommendation is to use a selection of measures (3-6 different) and use the patterns of generated groups that are similar across the different distance measures.

Regarding which variable(s) to use for the distance measure, the situation is tricky. For each of the groups a model is estimated and a set of coefficients produced. Groups with similar coefficient estimates are then grouped together; whereafter common characteristics of these collections of groups are collected in order to form the descriptions of the segments. To describe the common characteristics, it is necessary that the observations within these collections not are too different – this would make the detection next to impossible. Thus to ease this, the observations in the initial groups should not be very different.

Stated differently; when groups are formed such that observations (within the groups), are similar with respect to a few variables, a segment description where observations (within the segments) are similar, is enabled. If observations within the groups generally are different with respect to some variable $Y$, a description of segments by levels of $Y$ above/ below some threshold value is not likely.

To resume, the list of variables used to form the groups may be the same as those used to describe the segments. However, they should not be in too much conflict with the list of variables expected to describe the segment. Further, if the segmented model is to be used for forecasts, it is necessary that the variables used to described the segment membership are also available at the forecast time (or are possible to forecast).

The recommendation is to use variables like income group, sex, age group, number of cars in household, possession of multi-ride ticket for bus/ train, number of children in household, etc. The common denominator for these variables is that they are descriptive of the individual and are expected to possibly influence the relative preferences, hence, the travel patterns of the individual. Often when data is collected, at least in Denmark, this information is collected, but not used for other purposes than statistics. Further, variables like length of the performed trip (5-8 classes), purpose of the performed trip, etc. which are summarised as person-trip specific characteristics.

A parallel between the group formation and the sample based techniques described in section 4.7 can be drawn. In the Bootstrapping method the parameters of interest are estimated on numerous re-samples drawn from the available data sample. One re-samples is drawn with replacement by random sampling, and the estimation is performed on the drawn re-sample, nothing is done with the remaining part of the data; a new re-sample is drawn by random sampling, with replacement, etc. In the SEP method groups are drawn without replace-
CHAPTER 5. DATA SEGMENTATION BY PREFERENCES (SEP METHOD)

ment, until all records have been drawn for a group. The groups are not drawn by random sampling – they are formed by cluster analysis using some previous knowledge (socioeconomic variables). Estimation is performed for each group. Only for the hierarchical segments, new re-samples (new sets of clusters) are generated; model estimation is performed for these too.

Similarly, it is attempted to draw a parallel to the Jackknife method. In the Jackknife one (group of) observation(s) is deleted from the sample to obtain the re-sample. The estimation is performed on the re-sample; a new re-sample is drawn by removing another (group of) observations, etc. The groups are of equal size, the collection of groups spans the data sample and groups are non-overlapping. When a group of observations are deleted this is 'without replacement' as every observation only can be removed once! The parallel to draw to the SEP method is the splitting up of the sample into groups, though the SEP make use the group \( X \) itself while the Jackknife uses the sample 'minus' the group \( X \setminus X^g \). In both cases the estimation results are weighted equally (enabled by the equal group size).

5.3.2 Identification of segments

Identification of segments requires two processes; the formation of segments and the description of members of each segment. At this point (step 4 in the algorithm) models have been estimated for each data group and stage is set for determining in which groups of observation the observed choice can be explained by the same behaviour or relative preferences.

The (relative) preferences for each attribute of travel, is here represented by the coefficient estimates for each group. It must be identified which groups are similar in preferences, as these form the basis for the segments. To illustrate the process a small example is given. The data is divided into 5 groups for which separate coefficient estimates are obtained. The model is based on three coefficients; group number, coefficient estimates, ratios of coefficients are given in table 5.1.

As the purpose is to find out which groups are described by similar preferences, a plot of the relative preferences \((\beta_3/\beta_1 \times \beta_2/\beta_1)\) is given in figure 5.4. The numbers 1-5 refer to group number. From the plot it can be seen that the relative preferences are similar for groups 1, 4 and 5; and is similar (at a different level) for groups 2 and 3. Hence, one cluster is formed of groups 4, 1 and 5 (cluster A) and one cluster is formed of groups 2 and 3 (cluster B).

This example was constructed with few groups and so few coefficients that plotting was an option for detecting the similarities. In a real case, 100-500 groups each described by 5-15 coefficient could be the case. Then detecting which groups are described by the same preferences by scatter plots, becomes a challenge! For this, cluster analysis (see section 4.4.2) can be used. The rel-
Table 5.1: Example of determination of cluster membership

<table>
<thead>
<tr>
<th>Group</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\beta_2/\beta_1$</th>
<th>$\beta_3/\beta_1$</th>
<th>$\beta_2/\beta_3$</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.00</td>
<td>-1.50</td>
<td>1.00</td>
<td>1.50</td>
<td>-1.00</td>
<td>-1.50</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>-0.70</td>
<td>-1.25</td>
<td>1.30</td>
<td>1.79</td>
<td>-1.86</td>
<td>-0.96</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>-0.80</td>
<td>-1.40</td>
<td>1.30</td>
<td>1.75</td>
<td>-1.63</td>
<td>-1.08</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>-1.00</td>
<td>-1.45</td>
<td>0.95</td>
<td>1.45</td>
<td>-0.95</td>
<td>-1.53</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>-1.00</td>
<td>-1.55</td>
<td>1.05</td>
<td>1.55</td>
<td>-1.05</td>
<td>-1.48</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 5.4: Example of determination of cluster membership. Note: numbers refer to groups.
relevant variables to minimise the distance between within the clusters, are the coefficient estimates; small distance is an analog to similarity. When the distance between the vector of group estimates is used, then the groups are similar in preferences in all the dimensions (i.e. their preferences towards travel time, cost, seat availability, etc.).

After the determination of which groups are in which clusters, the description of members (the observations) in each cluster follows. Description can be like *female, age 35+, 2+ cars in household or female, work based trip, low income* etc. The suggested way of doing it, is by trial and error in a modified way. Plot socioeconomic variables in a scatter plot, such that each point, which represents an observation, is labelled by cluster membership (1, 2, etc.). If a cluster is clearly confined to some area and observations in this area are (virtually) all members of this cluster, the segment is identified. However, it is not guaranteed that such a description exists! This is illustrated in figure 5.5. Members of cluster 1 are clearly confined to an area without members of other clusters. Hereby, cluster 1 is translated to segment 1 and is described by variable $1 \geq v_1$, variable $2 \geq v_2$. Clusters 2 and 3 are not clearly defined by combinations of these two variables, wherefore combinations of other variables are consulted.

However, the division of the areas with different cluster membership, may not always be as clear as in figure 5.5. If a few observations with ‘wrong’ cluster membership (2) are mingled into an area with a predominantly membership of another cluster (1), these observations are transferred to segment 1. This is exactly how the difference between clusters and segments occur.
5.4 Discussion

The description of the segments is by far limited to those socioeconomic variables used in the model - in some cases it might prove advantageous to use other variables, eg. gender, age, number of children in household etc. Moreover, it is worthwhile to keep in mind that the segments are constructed from a data set (and not the whole population) wherefore use of these revealed segments in other respects requires further validation (tests for transferability).

Concentration of observations belonging to the same cluster is ensured to some extent. Members of each group are per construction, similar with respect to some variables (those used to form the groups), and members of each group are per construction, members of the same cluster.

5.3.3 Hierarchical segments

The algorithm may be extended to detect hierarchical segments, where each sub-segment exclusively belong to one segment. The method is to first identify clusters and within these clusters, sub-clusters are identified. Clusters are then 'translated' to segments whereafter the sub-clusters are 'translated' into sub-segments. Hierarchical segments are further described in section 4.3.2.

The algorithm may be extended with a loop over model estimation and cluster formation to improve clusters, and recognise sub-clusters. The algorithm is applied to form segments at the upper level. For each of these segments the algorithm is repeated to form sub-segments within each of the segments. The procedure can be repeated the same number of times as the required number of level of segments. This is a 'top-down' approach using the term from section 4.3.

Alternatively, hierarchical segments can also be formed by an additional round of cluster formation/segment identification of the segments obtained. This can be seen as the 'bottom-up' approach.

Hence, the formation of clusters can be based solely on the clearest division of data as well as a more hierarchical segmentation structure can be formed. This step wise cluster formation is illustrated in figure 5.6 where the dotted blue line indicate the initial division into four clusters – these are subsequently aggregated into two clusters (indicated by symbol). This may prove advantageous in strategic planning of a market. First, the relevant market is pinpointed, then sub markets may be addressed separately.

5.4 Discussion

Traditional segmentation methods have been based on threshold values of one (few) variables in the data set and not based on latent characteristics of the individuals in the sample (segmentation by purpose e.g. (Nielsen and Jovicic, 1999; Nielsen et al., 2000), trip length, age and employment status (Kulkarni
and McNally, 2000)). The main problem with segmenting by latent variables is indeed that they are unobservable. The proposed method seeks to work around this problem, though retaining itself to stay within the limitations of how many model coefficients may be estimated from a data set (four coefficients can not be estimated from two data records, as the model is not identified).

Any method based on mathematics and/or statistics has limitations and implicit assumptions – so does this method. As of first, any assumption of the discrete choice model is in effect. That is for e.g. the MNL model, independence of alternatives, independent and identical distributed individuals (IID residual term). Implicit assumptions are that the data may actually be described by the model type and specification of the utility function. But for a segmented model, the assumption on the utility function only apply for data within segments, not across the whole sample. In other words, the assumption is less restrictive for the segmented model, than for the un-segmented model.

Less obvious is the assumption that segments may be described by variables in data set. Given that that the data may be segmented, and that these may – in some way – be described by some variable in the data, it makes sense to search for the segments; otherwise not. The author is not aware of the existence of a sophisticated method to test for existence of segmentation, apart from ‘trial-and-error’ based on improved model fit\(^9\).

\(^9\)In practical application, the model fit of the segmented model is compared to the un-segmented model via F-test or likelihood ratio test, taking into account the larger number of coefficients in the segmented model. Though, often what is not taken into account is that successive test of model specifications on the same data is bound to at some time, imply significance of an in fact non-significant model (known as Type II error) or reject a significant model (Type I error).
5.4 Discussion

Given that data may be segmented and clusters are found that improves the likelihood of the overall model, identification of the segments is still an issue. Identification or the 'translation' of clusters into segments requires some judgement as to which attributes may possibly define these segments (gross list) and of these, how segments actually are described, e.g. threshold values of (functions of) attributes. The list of describing variables may be composed of attributes in the utility function but may also consist of other socioeconomic variables as e.g. number of cars in the household, age potentially combined with gender, etc.

Repetition of the model (see section 5.3.3) stand for forming segment that include sub-segments, where later repetitions form segments within the previously formed segments. This construction may prove valuable if a new marketing strategy is under consideration as the (sub) segments may be targeted differently, and the effect of this on traffic demand may actually be assessed, without doing the 'full scale' experiment.

As described earlier, a clear limitation of the method is the implicit assumption, that segmentation of the data is possible in the sense that is improves the model fit and degree of explanation. A search for a data partition based on latent variables (individual preferences) that will give rise to different estimated coefficients for each segment, resembles the problem of whether the chicken or the egg existed first. Estimates of coefficients are required to form the clusters though the clusters will be dependent on the estimates. On the other hand, given that segmentation is possible and will improve the model, the division into clusters (which records are members of which clusters) is required to estimate the model. Similarly for the segmented model, where knowledge of membership of the segments is required to estimate the segmented model.

Future work within data segmentation, includes existence of segments, as this is presently left unaccounted for. Further, it may be argued that segmentation is simply a special case of the Mixed Logit model (see section 3.6.2), where the mixing distribution is degenerate in each of the segments coefficient values. This in turn, suggests the search of an 'optimal' data segmentation by the ML model, where a multi-peaked distribution is employed and the number of peaks for the distribution is to be estimated within the model estimation. The distribution should be degenerate if identical preferences within each segment are assumed or continuous if a Mixed Logit specification is assumed for each segment. Though, if the model is to be used for forecasting, it may prove advantageous to translate the distribution into segments (where possible) to avoid problems of extrapolating the shape (and parameters) of the distribution for use in forecasting.
5.5 Empirical evidence

The SEP method has been tested on one data set. The data was collection for the Copenhagen-Ringsted model, a traffic model covering the eastern part of Denmark. When the Copenhagen-Ringsted model was constructed (1998-99), the model was formed based on the available data, re-using existing model code and existing segmentation of data. Individuals were assigned to segments depending on the purpose of their trip, as it has been recognised that individuals travelling with the purpose ‘to/from work’ generally could be described by a set of relative preferences different from an individual doing a shopping trip. Models formed on un-segmented data as well as for each of the purpose-defined segments, were compared in the light of model fit (likelihood) and coefficient values (based on sound knowledge).

Subsequently, the proposed SEP method for data segmentation was applied to the RP data from the Copenhagen-Ringsted model. The data consisted of 3 different RP data sets, a total of 6700 records. In this case each observation corresponded to one individual performing one trip and there were only one observation for each individual, thereby no potential for repeated measurements error to account for. For a description of the data and/or the model see Nielsen (2000).

Six different model specifications were estimated; (1) an unsegmented model for reference, two purpose-segmented models – (2) based on three segments, (3) based on five segments. Finally, three based on the SEP method. The model requires initial division of data into groups. The difference between three models is the number of groups, where model (4) is based on 50 groups while (5) and (6) are based on 100 groups. The segments produced by model (6) is a concentrate of the segments in model (5).

In further detail, the tests are;

1. A model for reference, which is based on unsegmented data. The model formulation has been tested during the formation of the Copenhagen-Ringsted model complex, data validation and utility functions are directly reused from there.

2. A purpose segmented model. Three segments are formed; home-work/education, Business and Other. Other model specifications as model 1.

3. A purpose segmented model. Five segments are formed; a further subdivision of the home-work/education and the ‘other’ segments resulted in five purposes. Other model specifications as model 1.

4. Application of SEP on raw data. Data is initially divided into 50 groups, 7 segments are formed. Other model specifications as model 1.
5.5 Empirical evidence

<table>
<thead>
<tr>
<th>Test</th>
<th>Log likelihood</th>
<th>Description</th>
<th># Coeff.</th>
<th>Segments identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-49,063.71</td>
<td>Unsegmented</td>
<td>5</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>-41,600.42</td>
<td>5 purposes</td>
<td>25</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>-43,308.87</td>
<td>3 purposes</td>
<td>15</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>-42,760.51 *</td>
<td>7 segments based on 50 groups</td>
<td>35</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>-40,559.39 *</td>
<td>10 segments based on 100 groups</td>
<td>50</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>-48,194.40 *</td>
<td>3 segments based on 100 groups</td>
<td>15</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 5.2: Results from application of the SEP method. * Log likelihood is for the cluster-based model.

5. Application of SEP on raw data. Data is initially divided into 100 groups, 10 segments formed. Other model specifications as model 1.

6. Application of SEP on raw data. 3 segments formed based on the 7 segments formed in model 5. This is an example of hierarchical segments. Other model specifications as model 1.

Unfortunately, the test was only conducted on a small data set in the sense, that only few variables were available in addition to those applied for the utility function. Clusters were formed and models based on the cluster-partitioned data were estimated.

The SEP method was applied to the data an the results are shown in table 5.2. In the table the log-likelihood value and the number of required coefficients to be estimated for each model is given. The last column indicate whether it was possible to unambiguously describe all the segments by the available data.

The results clearly show that the SEP method improves the log-likelihood of the estimated models. Furthermore, that model 5 is best in terms of likelihood value. However, in neither of the three models where the SEP method was applied it was possible to describe all the segments from the clusters.

Since the segments could not be described by the available variables, the case could not be used to determine the worth of the method. Further results from the application are in Appendix B.
5.5.1 Summary of empirical evidence

Summarised, the application of the SEP method to the data showed a potential for improving model fit in terms of likelihood values and coefficient values (based on sound knowledge). The test revealed a potential for improvement of the fit of model to data, compared to both an un-segmented model and a purpose-segmented model. Three models with alternative segments were constructed - two of these were significantly better than the purpose-segmented model. For this data set, a shortfall of the method was illuminated. The segments were constructed by means of cluster analysis (based on the coefficient estimates), the last step of the method is a 'translation' of the clusters into segments, that is, a description of the individuals (records) assigned to each segment. Whether this translation is possible or not, depends on the data as segments must be described by variables (socioeconomic, individual specific, etc.) in the data.

In this case, it was not possible to recover a simple description for all segments based on the variables in the data set (as only very few variables were available). Hereby, the case was 'too small' to determine the worth of the method. Specific results from the application are in Appendix B.

At this point, the only conclusion to be reached is that there is a potential for improving model fit by sub-division of data, but whether this can be operationalised by the SEP method remains an open question. Further tests on another data set, which includes more variables should be able to shed light on this issue.
Chapter 6

Assessment of shape of distributions (SODA Method)

6.1 Introduction and background

Models including error components are today at the frontier of application and development in transport modelling. The name Error Component Models are often used indiscriminately with Random Coefficients (Parameters) Logit, Models with Stochastic (Distributed) Coefficients (Preferences), Mixed Logit Models, Logit Kernel Models or lately Hybrid Choice Models, for models where additional stochastic terms are included in the traditional (linear) utility function.

At present the use of such models is growing rapidly, due to an increased access to software (both commercially by Hague Consulting Group (ALOGIT4ec) and freeware as GAUSS code at http://elsa.berkeley.edu/~train/software.html or the self-contained BioGeme at http://rosowww.epfl.ch/mbi/biogeme). In general, the method of Maximum Simulated Likelihood (MSL) is applied, described in section 4.6.1. However, the estimation by MSL only finds the optimum parameters for a given a priori distribution of the distributed terms. In comparison with estimation of non-mixed models, which are conditional on the specification of the utility function (which variables are included, linear or non-linear form), the estimation of the mixed models is further conditional on the a priori distribution of the distributed term. Only few of the analyses so far, has dealt with the interesting question of correlation between these error components, or the a priori assumption of type of distribution.

The method of simulated likelihood (MSL) is generally applied for estimation of mixed logit models, though alternatives exists, e.g. the Method of Simu-
CHAPTER 6. ASSESSMENT OF SHAPE OF DISTRIBUTIONS
(SODA METHOD)

lated Moments (see section 4.6.2) and Train (2003). Validation studies of the
MSL have shown nice performance regarding convergence and reproduction of
estimates. The current issue is of generation of random numbers used in the
estimation, to speed up estimation.

The question of which a priori distribution to apply for the coefficients, has
been somewhat neglected. This sections concerns this matter and proposes a
new method for assessing the (empirical) shape of distribution.

The urge for performing a MSL estimation conditional on the correct distri-
bution, was emphasised by a test on synthetic data. Several data sets were
constructed, of these five were constructed with a 'Random Parameter' specifi-
cation, that is the utility was described as \( U = \sum (\beta + \xi)X + \epsilon \), where \( \beta, X, \epsilon \) are
the traditional (fixed) coefficient, the explanatory data and the residual term,
while \( \xi \) is the distributed term added.\(^1\) The data sets differed by having
assumed different – though known, shapes of distribution.

Mixed models with correct parametrisation but wrong specification of distri-
bution, were estimated by means of MSL. These tests revealed that the MSL
estimation did not reject models based on a wrong distribution of the distributed
terms. On the contrary, the indication from the \( t \)-test was that distributed terms
(wrong distribution) did contribute to explain the variation in the data. This
finding even holds when a skewed distribution was estimated for a symmetric
distribution and the reverse! The case study is further described in section 7.1.

This point adds to the importance of finding the correct shape of distribution
by other means than a combination of MSL and \( t \)-test.

An alternative method to determine the distributions was proposed during the
Ph.D.-study and described in the papers (Sørensen and Nielsen, 2001, 2002;
Sørensen, 2002). This chapter describes the method in greater detail followed
by a chapter on applications. Section 7.1 contains applications to synthetic data
(testing the method) followed by applications to Danish mode choice data in
section 7.2.

In the following the method is referred to as the SODA method (Shape Of
Distribution Assessment).

6.2 SODA method for assessment of shape of
distribution

As mentioned above the MSL method estimate parameters of the distribution of
coefficients (or other distributed terms) in the utility function. The parameter

\(^1\)The Random Parameter Logit model is described in section 3.6.1
6.2 SODA method for assessment of shape of distribution

estimates are obtained conditional on the assumed distribution of these terms, where the alternative to the a priori assumed distribution is ‘no distribution’, in the sense that the variance term is estimated as ‘not significantly different from zero’. The objective was to enable an assessment of the shape of distribution, preferably assumption free – or at least without assumptions on which shape of distribution. Regrettably, the assumption of parametrisation must be maintained in order to make use of the probability model; similarly for any other assumption that relates to the probability model.

The purpose of the proposed method is to obtain an assessment of the shape of the possible underlying distribution of distributed terms – where there is no assumption on shape of distribution! The method seeks to uncover the empirical distribution of the data by repeated model estimations within a discrete choice framework. Preferably, we would like to estimate a model for each data record and compare the estimates for all these models to assess the shape of the distribution. The estimations are performed on smaller subsets of data since it is not possible to perform at model estimation for each individual. Each model estimation requires more data records than the number of coefficients to be estimated; this is a matter of identifiability of the models, see section 2.3.2. Generally a data set with multiple records per individual is not available wherefore the choice here has been to estimate models based on sub-samples (or groups of records) of the data.

The distribution of the coefficients, was assessed by numerous repeated model runs based on subsets of data based on a fixed model setup. Subsets (groups) of the original data was generated with respect to producing the largest number of groups possible, given that model estimation should be possible. Groups were obtained by random sampling aiming at equally sized groups allowing for the highest number of groups. Groups of data were disjoint and collectively exclusive of the data. For each group a demand model was estimated and the distribution of the coefficients was accessed from the collection of estimates (a $\beta$ vector for each group) from these model runs.

Determination of the shape of the distribution was the main purpose, that is, whether the empirical distribution is uni-modal, bell shaped, skewed or any other systematic (describable) shape. The actual parameters of the distribution(s) can subsequently be determined by MSL estimation, conditional on the empirical distribution.

A combination of plotting and charting (histograms) and estimated moments (means, variances, skewness and kurtosis measures) was used for the identification of distribution. It was valuable to keep in mind the later application in forecast tools when pointing at a distribution, as some distributions are easily parameterised (e.g. the normal and lognormal) and easily simulated whereas others may be quite burdensome e.g. draws from a 2 dimensional (correlated) distribution where each dimension follows different distributions.
However, the method is not suggested as an alternative to MSL, rather a complimentary method by which, the shape of distribution is determined prior to the MSL estimation. Hereby, the MSL estimation is based on the correct distribution and not, as is presently the case, some *a priori* assumption of shape of distribution.

### 6.2.1 Motivation for SODA method

The SODA method is based on statistical assumptions of the estimation procedure (likelihood theory) and random sampling theory.

In a traditional discrete choice model a probability model (e.g. MNL, NL) is assumed and estimated. For the SODA method a fixed-coefficient probability model e.g. MNL or NL, is assumed and estimated, for each group. For all groups the same model specification is applied, that is, the same type of probability model and the same specification of the utility function; that is, which variable to enter and the functional form (e.g. linear).

The data set $X$ consists of $N$ records each labelled $X_1, \ldots, X_N$. Let $G$ be the number of groups. Division of data $X$ into smaller groups (sup-samples) $X_g$ by random sampling preserves the distribution of the coefficients, i.e.

$$distr[\hat{\theta}(X_G)] = distr[\hat{\theta}(X)],$$

where $\hat{\theta}$ is an estimator.

### 6.3 Algorithm

To formalise the method an algorithm for the assessment of distributions may be set up as below. Each of the steps are discussed below with a general discussion of the method followed by details on the grouping procedure and determination of distribution in sections 6.3.1 through 6.3.3.

The initialisation (step 1) serves to set the stage. One of the keys to the SODA method is the division of the data sample into groups; the number of groups $G$ is inversely related to the number of observations in each group. The model estimation requires some minimum number of observations before the estimators for the coefficients are identified (described in section 2.3.2). Though, at the same time the number of groups should be as large as possible to enable the model estimation to detect possible differences between observations (larger groups tend to pull the coefficient estimates towards the estimates for the model based on the whole sample).

A recommended level for number of repetitions $R$ is 10 to 15 (for 100-500 groups); this serves to smooth the revealed distribution. If the number of groups is higher the number of repetitions can be lower. The recommended number of different group sizes $T$ is 2-4; is serves to determine whether the revealed distri-
### 6.3 Algorithm

**Step 1**  
Initialisation: Determine number of groups $G$ (or elements in each group, as these figures are inversely related), number of repetitions (fixed $G$) $R$ and number of different group sizes $T$.

**Step 2**  
Generate random groups.

**Step 3**  
Estimate models (a vector of coefficients) for each group.

**Step 4**  
Determine (simultaneous) distribution of coefficients.

**Step 5**  
Repeat steps 2 through 4, $R$ times.

**Step 6**  
Repeat steps 1 through 5, $T$ times.

**Step 7**  
Determine (simultaneous) distribution of coefficients based on pooled data.

After initialisation, step 2 proceeds with grouping (data partitioning) of the data. This partitioning is performed by random sampling to preserve the properties of the data. Other means of data partitioning (cluster analysis, proportional to some variable) were not recommended to use as they would make the consecutive model estimation conditional on the criterion used for the data partitioning. At this point it should be noted that the initial groups should be neither too small nor too large. The first case would cause inability to estimate models for each of the groups while the latter would make the assumption of homogeneity within the groups become more critical. In general, the number of groups should be maximised with regards to that no group size is below some threshold. In practice, this is obtained by equal group sizes. This is further described in section 6.3.1.

In step 3, estimation of the coefficient for each group follows. The same probability model (e.g. MNL, NL) and the same specification of the utility function is applied for all groups. The list of probability models to chose from consists of those where the residual term $\epsilon$ is the only distributed term (the NL model or the MNL models described in section 3). If the $t$-statistics for some group indicate that a coefficient should be excluded, this is ignored, as this would lead to a larger error, than the lack of significance of one coefficient at this stage of the model will give rise to.

The models estimated are subject to a fixed (predefined) model set up, i.e. specification of the utility functions is fixed. In principle, any discrete choice model (based on random utility theory) may be applied for the estimation of the demand model. Though, when applying the nested logit the condition of
non-decreasing scale parameter (for the Gumbel distributed error term \( \epsilon \)) for lower nests may be violated\(^2\). This should be tested within the analysis, though the order of the nests must be consistent for all model runs to ensure comparability.

Based on the coefficient estimates histograms are produced for each of the coefficients, in **step 4**. The objective is to get an intuition on which shape applies to each of the coefficient distributions; shapes may differ for different coefficients. Based on the graphical evidence, the list of possible shapes can be narrowed, e.g. the shape is uni-modal, flat, thick tailed, etc. The determination of shape, is further aided by the calculation of the first four central moments; the mean, the variance, the skewness and the kurtosis.

Since the moments estimated at this point are of the *between-group* variation and not of the *between-observation* variation, it is not recommended to statistically test whether the moments are significantly different from 0. The shapes revealed at this stage are conditional on the number of groups produced and the actual generation of groups. Determination of shape and moment estimation are further described in sections 6.3.2 and 6.3.3.

\(^2\)This issue is further discussed in section 6.4.
6.3 Algorithm

The coefficient estimates and hereby the histogram of these is subject to the model setup, and therefore to the actual groups selected. Further, if the number of groups \( G \), is low, e.g. 20, the histogram of the estimated coefficients may seem rather coarse. To make up for this, the whole estimation procedure may be repeated to 'smooth' the distribution depicted in the charts. When charts are produced they may seem rather edged, which may be caused by the setup of the estimation. In step 5, the group generation (step 2) and model estimation (step 3) are repeated for a number of times, e.g. 10 times. Hereby 10 times \( G \) coefficients estimates are obtained. The pool of the coefficient estimates (\( 10 \times G \)) are then used to produce the histograms (all coefficient estimates are weighted equally). Hereby the histograms becomes 'smoother' and the impact of the specific grouping is relaxed; this step is later referred to as the 'smoothing procedure' and is further described in section 6.3.2. Also, different groupings should be tested to increase robustness.

Similarly, the decision on number of groups (in step 1) may be crucial for the shape of distributions revealed. For the sake of investigating this step 6 was added. This step perform a loop over steps 1 through 5, and thereby shapes conditional on the number of groups are produced. The shapes have been similar over the different number of groups (\( G \)) tested (for the tested data sets), whereby the assumption of a specific number of groups has not appeared crucial.

Finally, in step 7 all the coefficient estimates are pooled and a search for distribution is conducted (identical to that in step 4). The pooling of data implicitly assigns an equal weight to all the coefficient estimates.

The method is illustrated in figure 6.1 working from left to right.

6.3.1 Data grouping

A key issue of the method is the data grouping. Some systematic approaches suggest themselves such as cluster analysis, partitioning proportional to some variable as well as the method of random sampling. The first two include some information in the groups which the following estimation is subject to.\(^3\) Random groups, on the other hand, preserves some of the statistical properties of the data e.g. the mean value. In addition, they are quite easily constructed.

By grouping the data a division of data into \( K \) smaller groups which are mutually exclusive (\( X^k \cap X^j = \emptyset, j \neq k \)) and totally exhaustive (\( X = \bigcup_k X^k \)) is referred to. Separate vectors of coefficients can be estimated for each group. Random groups can be obtained by use of a random sampler like the Gibbs sampler or any other variant of the Hastings-Metropolis algorithm\(^4\). A simpler alternative is to generate random group numbers by use the integer part of a

\(^3\)For a description on cluster analysis, see section 4.4.2.

\(^4\)For a description of random sample generation, see e.g. Ross (1997).
uniformly distributed number; that is, group number is $\lceil U(0, K) \rceil + 1$. The
generated groups are of fairly equal size, which is sufficient for this analysis. When data is constituted of more than one data set, an even spread between
groups is assured for each data set.

As the revealed distributions are conditional on the specific groups formed,
different groupings should be tested to minimise the distortion caused by how
a set of groups was selected (which observation was assigned to which group).
Further, the number of observations in each group may have an impact on
the estimation, hence on the distribution revealed. The counterpart of size of
groups, namely the number of the groups has an impact on the coarseness of the
frequency plot – this is described in the following section. The effect of these
may be reduced be generating different groups (same number of groups) [step
5] and altering the number of groups (and again generating different groups for
the ‘new’ number of groups) [step 6].

The formation of groups is done without replacement. For each chart produced
of empirical distribution, each and every data record contribute ‘one time’, in
the sense that each data record contribute to produce model estimates for ex-
actly one group. If the ‘smoothing procedure’ is applied, each record contribute
to exactly as many group-estimates as there are repetitions of the procedure.

A parallel can be drawn to the sample based techniques described in section
4.7. In the Bootstrapping method the parameters of interest are estimated on
numerous re-samples drawn from the available data sample. One re-samples
is drawn with replacement and the estimation is performed on the drawn re-
sample, nothing is done with the remaining part of the data; a new re-sample
is drawn with replacement, etc. In the SODA method groups are drawn
without replacement, until all records have been drawn for a group. Estimation is
performed for each of these groups; groups are drawn without replacement, etc.
The common characteristic is the repeated draw of re-samples or groups.

A parallel is harder to draw to the Jackknife method, where one (group of)
observations are deleted from the sample to obtain the re-samples. The esti-
mation is performed on the re-sample; a new re-sample is drawn by removing
another (group of) observations, etc. When a group of observations are deleted
this is ‘without replacement’ as every observation only can be removed once!
The flimsy parallel to draw, is the use of groups, although in the SODA method
the group, $X^g$ iteself is used and not the remainder, $X \setminus X^g$.

6.3.2 Shape of distribution
A second and important step in the algorithm is to characterise or preferably
identify the distribution. In the statistics literature numerous known distribu-
tions are described, but this is far from a complete listing. Some empirical
distributions have ‘never been seen before’ or rather, description of them have
never before been attempted. An 'accurate' description of a distribution may require a number of parameters – each of these parameters must be estimated from the data whereby the demand on data requirements rise. On the contrary, a description based on a large number of parameters is more sensitive to the data it is estimated from (some of the parameters describe minor variations in the shape). Further, an 'accurate' description of a distribution requires the same number of parameters to be forecasted if the model is to be used for forecast, recall the discussion of forecast in section 4.8. This in itself, is a very weighty argument for ending up with a simpler description of the distribution.

For a set of data two density plots are readily available; the mass density plot and the cumulative density plot. To identify the distribution from a plot of the cumulative density requires great skills! However, to identify or characterise a distribution from a density plot is not that sophisticated. To identify empirical distributions which are outcomes from a heuristic, the first stage is to broadly characterise the shape. Such a characterisation could include the some of the terms explained in table 6.1.

Having narrowed the shape, two things should be undertaken. First moments should be estimated to assess whether the immediate impression of shape is supported by data; estimation of moments are described in section 6.3.3. If it is believed that the shape of the distribution is 'close' to some well-known distribution, statistical tests like e.g. the Kolmogorov-Smirnov, Shapiro-Wilk, Anderson-Darling can determine whether the data is a realisation of some specified distribution. However, the data can only be compared with a pre-specified statistical distribution. These tests are briefly described in appendix A.4. The idea behind these tests is to compare the (relative) number of occurrences in some intervals with the probability for being in that interval for the particular distribution. Tests for whether the shape is similar to the normal distribution, can further be used to test for transforms of the normal e.g. the log-normal, by transforming the elements in the sequence.

Retrieved empirical distributions may seem rather coarse or edgy. This is due to the fact, that the frequency plot is only based on e.g. 50 coefficients estimates (the number of groups the data was partitioned into) but spread out on e.g. 10-15 intervals to get an impression of the spread. Each of the group estimates are weighted by $1/G$, the reciprocal of the number of groups, which is legitimated by the 'equal group size' aimed at during the formation of the groups (had the groups been of different size, the coefficient estimates should be weighted differently).

Repetitions of the procedure say $R$ times (as suggested in step 5), produces new vectors of coefficient estimates, based on the same assumptions of number of groups to be formed and the same model specification. Pooling of estimates results in $G \times R$ estimates for each of the coefficients for which new improved plots can be based. This is termed the 'smoothing procedure'. The weight for
CHAPTER 6. ASSESSMENT OF SHAPE OF DISTRIBUTIONS
(SODA METHOD)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Symptom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sign</td>
<td>Is the distribution restricted to either side of 0, the vast majority on one side or 'evenly' spread on the two sides?</td>
</tr>
<tr>
<td>Variation</td>
<td>Is the density mass spread out? 'evenly' on either side of the mean value?</td>
</tr>
<tr>
<td>Degenerate</td>
<td>Is the density mass concentrated in one (few) points? Or continuous along the axis?</td>
</tr>
<tr>
<td>Discrete</td>
<td>Is the density mass restricted to only having mass at distinct points along the axis?</td>
</tr>
<tr>
<td>Flat</td>
<td>Is the density mass spread out, with no substantial points of concentrations</td>
</tr>
<tr>
<td>Symmetric</td>
<td>Is the density mass symmetric around the middle mode (between the modes when number of modes is even)?</td>
</tr>
<tr>
<td>Bell-shape</td>
<td>Is the density mass uni-modal, symmetrical and the shape of the middle part of the density resembles a bell?</td>
</tr>
<tr>
<td>Uni-modal</td>
<td>Is the density spread out, with exactly one mode (point of concentration of density)?</td>
</tr>
<tr>
<td>Multi-mode</td>
<td>Is the density spread out, with more modes (local concentrations of density)?</td>
</tr>
<tr>
<td>Skewness</td>
<td>Is there more density mass at one side of the mode?</td>
</tr>
<tr>
<td>Thick tail</td>
<td>Is a substantial part of the density mass 'far' from the mean?</td>
</tr>
</tbody>
</table>

If distribution is allowed for more than one term:

| Correlation | Are the positive/ negative correlation between the terms? |

Table 6.1: Terms to describe the shape of a distribution
6.3 Algorithm

each of the coefficient estimates has now been reduced to $1/RG$; each single data record contribute to the estimation of $R$ estimates. Prior to pooling the data, the similarities of shape between the $R$ repetitions should be validated.

6.3.3 Moments and covariance

Distributions can be described by different means; from the accuracy of the distribution function if the distribution is known, over more waffling 'two-mode', uni-modal to the distribution free measures, moments. In order to aid in assessing the shape of the distribution, the mean value, variation as well as measures of skewness and kurtosis can be calculated (the first 4 moments), provided that the distribution is uni-modal (otherwise these moments do not contain much information).

The mean and variance needs no further introduction. These are also referred to as the first and second central moments, given by $\hat{m}_t = \sum_{j=1}^{N} (X_j - \bar{X})^t / N$, where $t$ is the order of the moment.

The third and fourth moments are here introduced as if the distribution is uni-modal and non-degenerate (bell shaped). Skewness measures the symmetry of the distribution. A skewness of zero indicates that values distribute evenly on both sides of the mean. A negatively (positively) skewed distribution, indicate that more values occur above (below) the mean, fewer are below the mean. The traditional (Pearson) measure of skewness is applied here, which is given by the third moment divided by the cubic of the spread $\hat{b}_1 = \hat{m}_3 / (\hat{m}_2)^{3/2}$ ($m_t$ is the $t$-th central moment) where $\sqrt{\hat{b}_1}$ is approximately distributed $N(0, \sqrt{6/n})$.

Although, several other measures of skewness exists (for a brief summary of measures of Skewness see Kotz and Johnson (1982-99)).

Kurtosis measures the heaviness of the tails as compared to the normal distribution. Peaked distributions, where values tend to concentrate in the middle, have a positive kurtosis. A negative kurtosis indicates a fairly flat distribution with heavy tails. Kurtosis is also referred to as a measure for 'heavy tails' compared to the normal distribution. The measure of kurtosis is given by the fourth moment over the square of the second moment, $\hat{b}_2 = \hat{m}_4 / (\hat{m}_2)^2$. Normality corresponds to $b_2 = 3$, values of $b_2$ above 3 indicates heavier tail than the normal distribution. The estimator of kurtosis employed ($\hat{g}_2$), measures the departure from normality, is $\hat{g}_2 = (\hat{m}_4 / \hat{m}_2^2) - 3$, distributed $N(0, \sqrt{24/n})$.

Some caution should be exerted when interpreting the moments. The higher the order of the moment, the greater the influence of outliers on the estimated moment, as the moments are given by the departure from the mean raised to the order of the moment. Or stated differently, if there are outliers in the data these

\[ 5\hat{g}_2 \geq -2, \text{ at all times, which can be proven by Cauchy-Schwarz’ inequality as } m_4 \geq m_2^2, \text{ thus } b_2 \geq 1 \text{ (Kendall et al., 1987).} \]
may distort the higher order moments quite heavily. For that reason, it does not make sense to perform a statistical test of moments significantly different from zero for the SODA method (as there are generally outliers).

If the analysis suggests a skewed distribution, some number of (renowned) distribution suggests themselves. Examples are the lognormal, inverse gauss, exponential and the gamma distributions. These are characterised by relatively simple formulae’s for the distribution function, though the lognormal has the advantage of easily producing draws from a simultaneous distribution. When applying the model for forecasts, simulated values drawn from the distribution is needed, whether the distribution is a (some) one-dimensional or a multi-dimensional simultaneous distribution.

The task of describing a distribution is by no means a simple task. Number of modes and characteristics of tails must be examined. Further, as more than one distribution is examined, possible interactions must be detected. Covariance measures the change in one variable given the change in the other; correlation is the corresponding measure on an invariant scale. Typically, by correlation we mean linear correlation measured from a sample (or a population) as

\[ r(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}. \]  

(6.1)

Significance of correlation can be tested by the simple test for (sample) correlation given by

\[ t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}, \]  

(6.2)

based on the \(N\) observations in the sample. The test statistic follows a Student’s distribution with \(N-2\) degrees of freedom. Confidence intervals can be determined by use of Fisher’s Z-transformation, which is generally used to test for hypotheses like \(r = r_0 \neq 0\). For a description of correlation measures see e.g. Spiegel (1998).

6.4 Discussion

The objective was to assess the shape of distribution to apply for demand modelling, for each of the coefficients that may involve a distribution (for those \(k\) where \(\xi_k \neq 0\)). The set-off for this method was the indulgence to assume simple distributional forms in recent MSL applications. This change from the fixed coefficients models’ inherent assumption of homogeneity among individuals where all have identical preferences, to acknowledging that individuals possess different preferences and to specify how they differ by choice of a priori distribution,

\(^6\)Another pursuit can be followed, which is describing the shape of the distribution by classification in Pearson’s system of moments, see e.g. Kotz and Johnson (1982-99).
is by and large a change from one assumption to another. Some recent examples are seen, where the specification of the utility functions allows for tests between up to three different distributions by using the relations between distributed terms (e.g. Nunes et al. (2001)). Other papers estimate models where different distributions are assumed, typically the normal and the lognormal distributions are assumed; examples are provided in section 3.6.1.

Such practice, does by the authors opinion, only give false confidence in the 'proven best' distribution, as this is proven better than a few fairly arbitrarily selected alternative distributions. But more significantly, comparisons to all other shapes of distributions are left unsettled! Hence, the distribution may be the best of say three or four chosen distributions, but may be less qualified than other distributions.

Determination of better or 'best' model, is typically based on significance of coefficients for explanatory variables and significance of variance for distributed terms. A test for significance of variance can be setup as a coefficient multiplied by a variable that follows a standard of the distribution in question, e.g. for a normal distributed term $\gamma \times \xi$ where $\xi \sim N(0, 1)$ is added, and $\gamma \neq 0$ is tested. Significance of a variance term ($t$-test of coefficient significantly different from zero) indicate whether this term contribute to explain some of the variation in the data, or not. This is not the same as significance of the distribution of that distributed term! An erroneously specified distributed term may explain some of the total variation, though a term distributed by the 'true' distribution may do so better.\footnote{This is discussed further in section 7.1.4, where the point is illustrated by use of synthetic data.}

Further, model fit e.g. the likelihood ratio test, can be used to determine whether a model including more coefficients is superior to a simpler model (see section 2.2.2).

This proposed method is to be seen as a tool to investigate distribution of coefficients and may be applied to RCL ($U^{RCL} = \sum_k X_k(\beta_k + \xi_k) + \epsilon$) and models specified with a random disturbance added to the utility ($U^{RD} = \sum_k X_k \beta_k + \gamma + \epsilon$). This method is not to be seen as an alternative to the MSL method for three main reasons. First, a crucial assumption is that individuals within a group, are assumed homogeneous (in order to make use of the e.g. MNL estimator) and this even for different groupings of individuals. Secondly, estimates of variation and the distributions obtained, are of between groups variation and not directly of between individual variation. Finally, the MSL is a generally recognised and well-proved method (with the shortfall of assuming the shape of the distribution), whereas the proposed method is a heuristic.

Regarding the assumption of identical preferences (homogeneity of individu-
als), the violation is equally or more severe for MNL models than for models based on grouped data, depending on how groups are formed. That is, in the MNL model, all individuals are assumed identical in preferences, whereas the assumption of identical preferences only applies to individuals within each group in the latter case. However, this assumption makes the model more restrictive than the ML model (where individuals are allowed to have different preferences).

The severity of the restriction is clearly depending on the size of the groups. A group size of one individual (either as one record or as several records for the same individual) is clearly not a restriction as long as the individual has consistent preferences over the multiple choice situations, though problematic as it cannot be estimated unless abundant data is available for each individual. Increasing the size of the group, facilitates estimation though at the cost of increased severeness of the assumption of identical individuals. Minimising the size of the groups (or likewise, maximising the number of groups), diminishes the error induced. To get a grasp of the impact, tests of different group sizes should be included when the algorithm is implemented, though implementations do not show a major problem.

To address the second issue of between group/between individual variation, a little more background on the maximum likelihood estimator is needed, see section 2.2.1. Given the fixed coefficients specification of the utility, it is assumed for the maximum likelihood estimation, that a single value of the coefficient describes the data. Previously, it has been recognised, that coefficient values do indeed vary over individuals. Estimates for each group are pulled towards some 'group value' leaving out some information on the variation between the individuals in the group. Hereby the span of coefficient estimates realised, is smaller than the actual span of 'true' coefficient values. The distribution depicted by this method is the distribution of these 'group values' and not of the whole sample, similarly for higher order moments.

If the 'true' value of the coefficients do indeed vary, it is likely that, by chance, some groups will contain a majority of 'high true value' individuals, whereby a higher than average estimate of the coefficient is realised. Or stated differently, groups leading to a high estimate of a coefficient, will reflect a predominantly high level of 'true' values for individuals in that particular group.

In the nested logit model, the order of the nests are determined by the size of the logit scale factor \( \mu \), such that scales are non-decreasing for lower nests.

---

8 Distributed coefficients may be subdivided into different individual behaving differently (systematic differences in relative preferences between individuals) versus the same individual behaving differently at different times (systematic differences in relative preferences between time periods). The second type of variation is then left unaccounted for.

9 This is a matter of identifiability of the model, which was further discussed in section 2.3.2.

10 Unless, for the very special and hypothetical case where all individuals within each groups (drawn by random sampling), have identical 'true' value; values may differ between groups.
Say we have a model with two levels described as choice of an $A$ alternative and choice of a $B$ alternative, where these may be ordered either choice of $A$ alternative at the upper level and choice of $B$ alternative at the lower level $(A,B)$, or the opposite $(B,A)$. The estimation of the model performed by maximum likelihood estimation, determines the coefficients $\beta$ and the ratio of the two scales $\mu_A/\mu_B$ (one scale must be normalised to enable estimation). The order is then determined as $(A,B)$ if $\mu_A \leq \mu_B$ and $(B,A)$ if $\mu_B \leq \mu_A$. That is, during the model estimation, the order of the nests is determined by data, as the most likely order of the nests.\(^{11}\)

As an example, say we have a data set where the 'true' ordering of the nests in the tree is $(A,B)$ for most data records, but for some (few) data records the 'true' ordering of the nests in the tree is $(B,A)$. This difference in 'true' order of nests is not revealed in a nested logit estimation. But, if by chance we take out a data group at random, which has a high representation of data records with a 'true' ordering of $(B,A)$, model estimation on this group give that $\mu_B \leq \mu_A$ and hereby that the order is $(B,A)$ on the contrary to what was the case for the model on the whole data. This would normally suggest that the order of the nests in the tree were reversed, but for the purpose of determining the distribution, this is neglected. Not that it is not a problem to violate assumptions of a model, but merely because the impact on the estimated coefficients by reversing the order for some (but not all) data groups is expected to be higher that the impact of violating the restriction of the relative size of the scales. This is not expected to happen frequently, and has not been detected in the data investigated so far.

The purpose of this method is to investigate further the shape of the distribution. To ease this search a 'smoothing procedure', which accumulates the distributional patterns for all model runs is incorporated before histograms, moments (1st - 4th) and correlation matrix are calculated. For data investigated so far, this has been sufficient for determination of shape of distribution.

The likelihood estimator of the coefficients is conditional of the specification of the model (MNL or other), the parameterisation of the model (construction of the utility function) and finally on the data it is estimated from (from different data sources, different coefficient estimates do generally arise). Coefficients are in general, interpreted as individual’s preferences for the various characteristics describing an alternative.

Given that it is plausible that coefficients are correlated, the focus of determining the distribution should be aimed at describing the majority of the variance by a simultaneous distribution rather than splitting the variance into (simpler) marginal distributions. The latter may lead to an incomplete description of the

\(^{11}\)The issue of ordering of nests in the nested logit model is further discussed in Ben-Akiva and Lerman (1985).
variation.

When applying the method to a model specification including alternative specific constants, one should consider whether distribution for these should be allowed or they should be fixed between models (and between runs). Nothing immediately suggests allowing for the distribution, as the constants are included in the model to patch up for a utility difference the deterministic part of the utility did not include. Furthermore, allowing for distribution of the ASC’s uses degrees of freedom in the specification preventing a large number of groups.

In addition to this, there are (at least) two different approaches to follow in determining the distribution. The simplest variant is to base it solely on the estimated coefficients, i.e. the vectors $\beta^k$ for each group $k$ from which the empirical distribution is easily depicted. A (probably) more appropriate path to follow is to incorporate the statistical variation of the coefficient estimates (confidence intervals or by quantiles) as well. However, in some cases this may prove to be a quite burdensome path to follow, and has not been investigated during this study.

If the determination of shape supports some particular well-described parametrised distribution, the parameters of this can often be calculated from the mean and the variance (two-parameter distributions). However, this is not recommended as the method for determining the parameters of the distribution, as the variation described in this manner is the between group variation and not the between observation variation which is strived for.

What is more important, is that as the (higher order) moments are very sensitive to outliers, the calculation of distribution parameters will also be highly influenced by outliers, if these are determined from the moments. Higher order moments are known to burst with increasing numbers of outliers. Estimation of distribution parameters by MSL is not nearly as sensitive to outliers as the SODA, since in the MSL estimation an 'averaging over draws' is performed. This averaging over draws implies an equal weight to each draw. MSL was described in section 4.6.1.

Applications of the method is given in chapter 7.
Chapter 7

Application

This chapter is on application of the two proposed methods to two different data sources. First, a section on synthetic data, where assumptions/limitations in the method for assessment of distribution (the SODA method) are tested. These results have not been reported before. This second section will revise a study on CRM data previously reported in Sørensen and Nielsen (2001). In the light of the section on synthetic data, distributions are revisited. Finally, the method for alternative segmentation (the SEP method) is applied to the CRM data.

7.1 Synthetic data

To test the method for assessment of shape of distribution six different synthetic data sets of 10,000 records were constructed. All constituted of choice between three alternatives, each described by a single attribute and two alternative specific constants (one removed from design for identification of model). An additional data set with choice described by two coefficients and two constant was constructed. Several different group sizes were tested, all between 50 and 500 groups (195 and 19 records per group, on average) to get a grasp of the effect of group size on the distribution.

To test the functionality of the method, different assumptions were incorporated in the synthetic data. As a base, the first data set includes one attribute multiplied by a fixed coefficient and no disturbances apart from the Gumbel error were added. The data set was intended as to give a grasp of variance incorporated by the method.

In the following three data sets (one attribute) random coefficients were added. Again, complexity increased from a uni-modal and symmetrical distribution (Normal), over a uni-modal but skewed distribution (LogNormal) to a bi-modal but symmetrical distribution. In the following data set, choice was generated by
adding a normal distributed disturbance to the (linear) utility (fixed coefficient for the attribute).

The final two (one-attribute) data sets differed from the before mentioned, as one attempted to replicate a data set collected by use of stated preference method, in the sense, that several records were created from the same preferences (coefficient values), though different for different 'individuals'. The coefficient employed for this set was based on the RCL specification with a Log Normal distribution.

Two attributes each multiplied by differently distributed coefficients (different shape and variation) described choice in the final data set. This set was formed to test whether the method can distinguish between the two distributions, can detect two different shapes and not least, detect the correct shape for each coefficient.

For each data set, the data generation is described in more detail in the following. A more compact notation is used for reference. The notation is [choice type][number of coefficients where distribution is allowed]. Choice type is F if choice is based on Fixed coefficients. Similarly, N, LN respective TP, if choice is based on 'Normal distributed RCL specification', 'Log Normal distributed RCL specification' respective 'Twin-Peaked distributed RCL specification'. ND is used for choice based on fixed coefficients and a Normal distributed added to the utility function. NLN is used for the two coefficient model with one normal and one lognormal distributed coefficient. Number of coefficients where distribution is allowed is either three or one, in the latter case the MNL estimated values for the alternative specific constants are used and only $\beta$ is estimated for each model (4 or 2 coefficients are estimated for the NLN model). As an example, $F3$ ($F1$) refers to the model based on fixed coefficients choice, where distribution is allowed for 3 (1) coefficients.

Comparisons of the data sets, are presented with respect to changes in chosen alternative and performance of the MNL estimation. Mixed Logit estimation results (by MSL) are presented where software limitations allows for so (just the normal and the lognormal distributions are included in the software). Finally, a presentation of model results where the SODA method has been applied.
7.1 Synthetic data

7.1.1 Description

Fixed coefficient, \( F \)

In this data set no distribution was added apart from the Gumbel error. 10,000 observations were generated by

\[
\begin{align*}
U_{1n} &= \beta X_{1n} + \alpha_1 + \epsilon_{1n} \\
U_{2n} &= \beta X_{2n} + \alpha_2 + \epsilon_{2n} \\
U_{3n} &= \beta X_{3n} + \epsilon_{3n}
\end{align*}
\] (7.1)

where \( \alpha = (1.5, 0.5) \), \( \beta = -1 \) and \( \mu = 1 \).\(^1\) Data \( X \) follows the \( N(10, 3^2) \) distribution\(^2\) and chosen alternative is the alternative with the highest utility. A constant term is left out in the \( U_{3n} \) for sake of identification of the specification.

This data set may give an idea of the variance of the estimation in the new method.

Normal RCL, \( N \)

A random disturbance (Normal distributed) was added to the coefficients. The same disturbance was added for each alternative, as this described an individual’s preference for a specific attribute, all else being equal.

\[
\begin{align*}
U_{1n} &= (\beta + \xi)X_{1n} + \alpha_1 + \epsilon_{1n} \\
U_{2n} &= (\beta + \xi)X_{2n} + \alpha_2 + \epsilon_{2n} \\
U_{3n} &= (\beta + \xi)X_{3n} + \epsilon_{3n}
\end{align*}
\] (7.2)

where \( \xi \) was normal distributed, \( N(0, 0.5^2) \). The remaining coefficients were as for data set \( F \).

Log Normal RCL, \( LN \)

A random disturbance (Log-Normal distributed) was added to the coefficients added. The same disturbance was added for each alternative, as this described an individual’s preference for a specific attribute, all else being equal.

\[
\begin{align*}
U_{1n} &= (\beta + \xi)X_{1n} + \alpha_1 + \epsilon_{1n} \\
U_{2n} &= (\beta + \xi)X_{2n} + \alpha_2 + \epsilon_{2n} \\
U_{3n} &= (\beta + \xi)X_{3n} + \epsilon_{3n}
\end{align*}
\] (7.3)

where \( \xi \) was log normal distributed, that is \( \log\xi \sim N(0, 0.5) = N(\zeta, \sigma^2) \) and \( E(\xi) = \exp(\zeta)\sqrt{\exp(\sigma^2)} = -1.13 \), whereby \( E(\beta + \xi) = -2.13 \).

\(^1\) Setting the logit scale \( \mu = 1 \) implies the mean is \( \eta + \gamma/\mu \approx 0.58 \) and variance of the error term of \( \pi^2/4\mu^2 \approx 1.65 \). These setting were applied for all synthetic data sets.

\(^2\) Records with negative values for either of \( X_1, X_2, X_3 \) are removed, as data typically employed for discrete choice models is sign specific, e.g. cost, travel time, level-of-service stated on a scale.
Twin-Peaked RCL, TP

A random disturbance (twin peaked) was added to the coefficients added. A bimodal distribution represents the case where the sample may be split into two segments each with their distribution. For this case these distributions have been specified with the same shape and variance but different mean values. The same disturbance is added to each alternative, as this describes an individual's preference for a specific attribute all else being equal.

\[ U_{1n} = (\beta + \xi)X_{1n} + \alpha_1 + \epsilon_{1n} \]
\[ U_{2n} = (\beta + \xi)X_{2n} + \alpha_2 + \epsilon_{2n} \]
\[ U_{3n} = (\beta + \xi)X_{3n} + \epsilon_{3n} \]

where \( \xi \) is simulated from two (different) triangular distribution on \([-0.5, 0]\) respective \([0, 0.5]\) such that each even (odd) record draws from the first (second) peak, the distribution has a mean 0. The distribution is shown in figure 7.1. In the following the distribution is termed 'Twin Peaked', for want of a better name.

Normal Disturbance, ND

Again, 10,000 observations were generated –in this case a normal distributed variate was added to the utility function, to assess the reaction of the method on data when data were not solely described by a logit model (residual term the sum of a Gumbel and a normal term). The 'probit with a logit kernel' would be a suitable model specification. Data was generated by

\[ U_{1n} = \beta X_{1n} + \sigma_1 \zeta_{1n} + \alpha_1 + \epsilon_{1n} \]
\[ U_{2n} = \beta X_{2n} + \sigma_2 \zeta_{2n} + \alpha_2 + \epsilon_{3n} \]
\[ U_{3n} = \beta X_{3n} + \sigma_3 \zeta_{3n} + \epsilon_{3n} \]

(7.5)
with $\alpha = (1.5, 0.5)$, $\beta = -1$, $\sigma = (3, 2, 1)$ and $\mu = 1$. Data $X$ follows the $N(10, 3^2)$ distribution and the utility (hence, the chosen alternative) of alternatives are calculated by use of a single draw of the disturbance $\zeta$ from a standard normal. A constant term is left out in the $U_{3n}$ for sake of identification.

The model to be estimated (tested for distribution) differs slightly from the above model, as it is specified as a random coefficients model (the above model is not identified due to the additivity of the normal distribution). The formulation is

\begin{align*}
U_{1n} &= \beta X_{1n} + \sigma_1 \zeta_{1n} + \alpha_1 + \epsilon_{1n} \\
U_{2n} &= \beta X_{2n} + \sigma_2 \zeta_{2n} + \alpha_2 + \epsilon_{2n} \\
U_{3n} &= \beta X_{3n} + \epsilon_{3n}
\end{align*}

Due to the difference in specification the utility function, the method is not expected to be able to recover the parameters of the distribution (the estimated distributional parameters are for the distribution of the differences of distributions and not the distribution itself).

**'Stated Preference' imitation, $SP$**

This data set is constructed as to mimic a Stated Preference data set, where each of 500 'respondents' performs 20 choices (index $t$) based on the same preferences, that is coefficients, a total of 10,000 records. Each group consists of just one 'respondent' hence the assumption of identical preferences within the group is not limiting. A random disturbance (log-normal distributed) was added to the coefficients added (the same used for each 'respondent’s subsequent choices).

\begin{align*}
U_{1n}^t &= (\beta + \xi) X_{1n} + \alpha_1 + \epsilon_{1n} \\
U_{2n}^t &= (\beta + \xi) X_{2n} + \alpha_2 + \epsilon_{2n} \\
U_{3n}^t &= (\beta + \xi) X_{3n} + \epsilon_{3n}, \forall t \in \{1, 2, \ldots, 20\}
\end{align*}

An important implication of this data type is that SODA repetitions are not possible, as the link between records and individuals is unique\(^3\). All later results for the $SP$ data is based on one model run, which is comparable to the average over repetitive model runs for the other data sets.

**Normal + Log Normal RCL, $NLN$**

Unlike the other data sets, in this data set two attributes ($X$ and $Y$) are included in the utility; both are multiplied by a stochastic coefficient. The first coefficient is assumed normal distributed, whereas the second is assumed lognormal. Mean values and variances are different for the two distributed terms, and they are

\(^3\)For the sake of model testing this may be overcome by using different seeds for generation of the random preferences - but as this does indeed not resemble reality, this is not pursued any further here.
assumed independent. The same coefficients were used for all alternatives for each individual.

The purpose of this data set is to assess whether the method can retrieve the shape of the two different distributions correctly. Again, records with wrong sign for the attributes (either $X$ or $Y$) or coefficients were removed, resulting in a total of 9,762 records. The data was generated by the following formula,

$$
U_{1n} = (\beta + \xi)X_{1n} + (\beta_2 + \xi_2)Y_{1n} + \alpha_1 + \epsilon_{1n}
$$
$$
U_{2n} = (\beta + \xi)X_{2n} + (\beta_2 + \xi_2)Y_{2n} + \alpha_2 + \epsilon_{2n}
$$
$$
U_{3n} = (\beta + \xi)X_{3n} + (\beta_2 + \xi_2)Y_{3n} + \epsilon_{3n}.
$$

(7.8)

The mean value for $\beta_2$ can be calculated from $\xi$ as $\log \xi \sim N(0,0.7) = N(\zeta, \sigma^2)$, where $E(\xi) = \exp(\zeta)\sqrt{\exp(\sigma^2)} = 1.42$ as , whereby $E(\beta_2 + \xi) = 2 + 1.42 = 3.42$ (the negative sign was captured by the attribute). The attributes were generated as $X \sim N(10,3^2), Y \sim N(7,2^2)$. The assumption of independence of $X, Y$ and $\beta, \beta_2$ is met by using different seeds for the random number generators.

### 7.1.2 Comparisons of synthetic data sets

Generation of different data sets was described in the proceeding section, at this point differences between the data sets and especially the implications for the choice of alternative, is addressed. As the actual choice is generally the only thing observed concerning the individuals’ behaviour, differences between individuals may be described solely by alterations of choice. Cross tabulations of choice arising from the different data sets, are shown below in tables 7.1 through 7.6. The number of records is 9,769\(^4\) after removal of observations with wrong sign for either of coefficient and attribute, for any of the data sets.

The major difference is observed between choice based on pure deterministic preferences, $F$ (though with Gumbel disturbance) and choice with a normal disturbance added, $ND$. As much as 26.59 % of the records have a change in choice (that is, the sum of off-diagonal elements).

As the data is simulated and subsequently some records have been removed due to sign change of the $\beta$-coefficient, estimated moments are listed below in table 7.7. As can be seen the moments (mean and variance) are close to the specified values.

### 7.1.3 MNL models of synthetic data sets

As in any other study, this study initialises by estimating MNL models for each of the data sources (choice formations). Results are presented below in table 7.8 where some interesting features occur. Keep in mind the value of the true

\[^4\text{For the two coefficient model, the number of records is 9,762.}\]
### 7.1 Synthetic data

<table>
<thead>
<tr>
<th></th>
<th>Choice by $ND$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>$\Sigma$</td>
</tr>
<tr>
<td>1</td>
<td>33.17</td>
<td>6.68</td>
<td>4.85</td>
<td>44.70</td>
</tr>
<tr>
<td>2</td>
<td>5.76</td>
<td>22.09</td>
<td>2.32</td>
<td>30.18</td>
</tr>
<tr>
<td>3</td>
<td>4.51</td>
<td>2.47</td>
<td>18.14</td>
<td>25.12</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>43.44</td>
<td>31.24</td>
<td>25.31</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Table 7.1: % changes of chosen alternative, from 'Fixed coefficients' to 'Normal Disturbance'

<table>
<thead>
<tr>
<th></th>
<th>Choice by $N$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>$\Sigma$</td>
</tr>
<tr>
<td>1</td>
<td>41.78</td>
<td>1.42</td>
<td>1.50</td>
<td>44.70</td>
</tr>
<tr>
<td>2</td>
<td>2.27</td>
<td>27.13</td>
<td>0.78</td>
<td>30.18</td>
</tr>
<tr>
<td>3</td>
<td>1.83</td>
<td>0.90</td>
<td>22.39</td>
<td>25.12</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>45.88</td>
<td>29.45</td>
<td>24.67</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Table 7.2: % changes for chosen alternative, from 'Fixed coefficients' to 'Normal RCL'

<table>
<thead>
<tr>
<th></th>
<th>Choice by $LN$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>$\Sigma$</td>
</tr>
<tr>
<td>1</td>
<td>36.73</td>
<td>3.79</td>
<td>4.19</td>
<td>44.70</td>
</tr>
<tr>
<td>2</td>
<td>0.97</td>
<td>27.02</td>
<td>2.18</td>
<td>30.18</td>
</tr>
<tr>
<td>3</td>
<td>0.59</td>
<td>1.13</td>
<td>23.40</td>
<td>25.12</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>38.29</td>
<td>31.94</td>
<td>29.77</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Table 7.3: % changes for chosen alternative, from 'Fixed coefficients' to 'Log-Normal RCL'

<table>
<thead>
<tr>
<th></th>
<th>Choice by $TP$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>$\Sigma$</td>
</tr>
<tr>
<td>1</td>
<td>43.82</td>
<td>0.40</td>
<td>0.51</td>
<td>44.73</td>
</tr>
<tr>
<td>2</td>
<td>0.48</td>
<td>29.40</td>
<td>0.24</td>
<td>30.12</td>
</tr>
<tr>
<td>3</td>
<td>0.61</td>
<td>0.34</td>
<td>24.20</td>
<td>25.15</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>44.91</td>
<td>30.14</td>
<td>24.95</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Table 7.4: % changes for chosen alternative, from 'Fixed coefficients' to 'Twin-Peaks RCL’
CHAPTER 7. APPLICATION

Table 7.5: % changes for chosen alternative, from ’Fixed coefficients’ to ’Log-normal RCL, SP’

<table>
<thead>
<tr>
<th></th>
<th>Choice by SP</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>Σ</td>
</tr>
<tr>
<td>1</td>
<td>38.02</td>
<td>3.24</td>
<td>3.47</td>
<td>44.73</td>
</tr>
<tr>
<td>2</td>
<td>0.81</td>
<td>27.56</td>
<td>1.75</td>
<td>30.12</td>
</tr>
<tr>
<td>3</td>
<td>0.51</td>
<td>1.01</td>
<td>23.63</td>
<td>25.15</td>
</tr>
<tr>
<td>Σ</td>
<td>39.34</td>
<td>31.81</td>
<td>28.85</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Table 7.6: % changes for chosen alternative, from ’Fixed coefficients’ to ’Normal + Log Normal RCL’

<table>
<thead>
<tr>
<th></th>
<th>Choice by NLN</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>Σ</td>
</tr>
<tr>
<td>1</td>
<td>24.26</td>
<td>11.99</td>
<td>8.80</td>
<td>45.05</td>
</tr>
<tr>
<td>2</td>
<td>8.15</td>
<td>18.34</td>
<td>3.88</td>
<td>30.37</td>
</tr>
<tr>
<td>3</td>
<td>6.68</td>
<td>8.41</td>
<td>9.49</td>
<td>24.58</td>
</tr>
<tr>
<td>Σ</td>
<td>39.09</td>
<td>38.74</td>
<td>22.17</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Table 7.7: Actual moments of simulated parameters

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>β_N</td>
<td>-1.02</td>
<td>-1.01</td>
<td>0.22</td>
<td>-0.20</td>
<td>-0.28</td>
</tr>
<tr>
<td>β_LN</td>
<td>-2.14</td>
<td>-2.01</td>
<td>0.34</td>
<td>-1.68</td>
<td>4.62</td>
</tr>
<tr>
<td>β_TP</td>
<td>-1.00</td>
<td>-0.99</td>
<td>0.02</td>
<td>0.00</td>
<td>-1.48</td>
</tr>
<tr>
<td>β_SP</td>
<td>-2.13</td>
<td>-1.99</td>
<td>0.34</td>
<td>-1.33</td>
<td>2.40</td>
</tr>
</tbody>
</table>

Table 7.7: Actual moments of simulated parameters
7.1 Synthetic data

parameters $\alpha = (1.5, 0.5)$, and $\beta$ as in table 7.7.

Two measures of mode fit are listed, the $\rho^2$ and the likelihood. The $\rho^2$ is 0.6 for most, even better 0.79 for the $LN$ and $SP$ model. The fit of the $ND$ model is substantially lower at 0.36, which is due to the MNL models inability to capture the random disturbance in the data. The likelihood value is more diverse with a range from -2260 to -6889, where the $LN$ model has the best fit, closely followed by the $SP$.\footnote{This is to be taken as a loose comparison, as likelihood is measured on an arbitrary scale. The value of these models is not comparable like likelihood ratio test, as they are not constructed as hierarchical specifications.} At this point it should be mentioned that the 'repeated measurement' has not been taken into account, this would improve the model fit. The third best is the $F$ which is rather surprising. Having no prior information of the coefficient values and relying solely on the test statistics ($\rho^2$ and $L$), the MNL model appears to be better a estimating the $LN$ model than the $F$! In the other end the $ND$ model stands out with the lowest likelihood value, which combined with a rather low $\rho^2$ indicates use of a wrong model specification for the data, which is indeed the case. The major difference is in the fit of the model to the data, where the MNL model seem to have the best fit to the data generated by lognormal distributed random coefficients and not as expected, the fixed coefficients data. The fit of the model to the data with a normal distributed random disturbance, is rather small compared to the other data sets, both in terms of likelihood value and $\rho^2$. This could suggest the use of a probit model – all though this information is generally not present in an applied modelling context.

Parameter recovery, as an indicator of model fit is only of value if the models actual scale is known as $\mu\beta$ and not $\beta$ is estimated. The importance of estimating the correct relative and absolute coefficients were discussed in section 4.8.

In this case where the 'true' coefficients are known, the ratio of the estimate to the 'true' may be used. For most models ($LN, TP, F, SP$) this ratio is around 1, though 0.71-0.83 for the $N$ and 0.50-0.66 for $ND$. For $LN, TP, F, SP$ parameter recovery and unit scaling is obtained, whereas for $N, ND$ parameter recovery is obtained given a scale of 0.75 respective 0.55. The $t$-statistics are generally high, and fairly the same (indicating an equally likelihood of obtaining another data set with the same parameters).\footnote{The correct interpretation of $t$ statistics is, that in the long run with data from many samples a 95% interval calculated from each sample will contain the 'true' parameter approximately 95% of the time. Further interpretations are given in (Congdon, 2001)}

One model (the $NLN$) was estimated based on data where choice was described by two attributes multiplied by a normal respective a lognormal distributed coefficient. The 'true' coefficient values are given below in table 7.9. The normal distributed coefficient is identical to the normal coefficient in the 1
attribute model, whereas the lognormal coefficient is different in mean and variance (and seed). The mean of the $\beta_{LN}$ in the random generation is calculated by $E(\beta_{LN}) = E(\beta_2) + E(\xi_2) = -2 + (-1.42) = -3.42$. The actual mean (the mean of the generated values) is -3.29, that is, these are rather close and the difference is in part due to the random generation and in part to that records with wrong sign of attributes or coefficients are removed. Also, the distribution of the lognormal coefficient is 'wider' than the normal coefficient, as the variation is 1.07 compared to the 0.22 for the normal. As for the above models, the MNL model was estimated for the two attribute model for sake of comparison. Results are in table 7.10 below. The model does not replicate the true coefficient values – the deviance is not explained by scaling as neither the ratios are perfectly recovered.

Table 7.8: MNL Results for all synthetic data sets

<table>
<thead>
<tr>
<th></th>
<th>$F$</th>
<th>$ND$</th>
<th>$SP$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>$\sigma$</td>
<td>$t$</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>1.456</td>
<td>0.045</td>
<td>32.3</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>0.468</td>
<td>0.043</td>
<td>10.8</td>
</tr>
<tr>
<td>$\beta$</td>
<td>-0.994</td>
<td>0.017</td>
<td>-59.5</td>
</tr>
<tr>
<td>$\mathcal{L}$</td>
<td>-4206</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho^2$</td>
<td>0.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obs</td>
<td>9,769</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.9: Actual moments of simulated parameters, 2 attributes model

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_N$</td>
<td>-1.02</td>
<td>-1.01</td>
<td>0.22</td>
<td>-0.20</td>
<td>-0.28</td>
</tr>
<tr>
<td>$\beta_{LN}$</td>
<td>-3.29</td>
<td>-3.00</td>
<td>1.07</td>
<td>-2.93</td>
<td>19.74</td>
</tr>
</tbody>
</table>
7.1 Synthetic data

<table>
<thead>
<tr>
<th></th>
<th>Est.</th>
<th>σ</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>α₁</td>
<td>1.182</td>
<td>0.0418</td>
<td>28.3</td>
</tr>
<tr>
<td>α₂</td>
<td>1.223</td>
<td>0.0419</td>
<td>29.2</td>
</tr>
<tr>
<td>β₁</td>
<td>-0.394</td>
<td>0.00829</td>
<td>-47.5</td>
</tr>
<tr>
<td>β₂</td>
<td>-1.060</td>
<td>0.0177</td>
<td>-59.8</td>
</tr>
<tr>
<td>L</td>
<td>-4935</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ρ²</td>
<td>0.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obs</td>
<td>9,762</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.10: MNL Results for the 2 attribute data sets

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th></th>
<th></th>
<th>LN</th>
<th></th>
<th></th>
<th>ND</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>std.e.</td>
<td>t</td>
<td>Est.</td>
<td>std.e.</td>
<td>t</td>
<td>Est.</td>
<td>std.e.</td>
</tr>
<tr>
<td>α₁</td>
<td>31.655</td>
<td>0.646</td>
<td>49.0</td>
<td>1.471</td>
<td>0.059</td>
<td>24.8</td>
<td>0.835</td>
<td>0.033</td>
</tr>
<tr>
<td>α₂</td>
<td>31.242</td>
<td>0.459</td>
<td>68.1</td>
<td>0.508</td>
<td>0.054</td>
<td>9.5</td>
<td>0.329</td>
<td>0.034</td>
</tr>
<tr>
<td>β</td>
<td>-5.620</td>
<td>0.292</td>
<td>19.2</td>
<td>-0.491</td>
<td>0.030</td>
<td>16.5</td>
<td>-0.502</td>
<td>0.008</td>
</tr>
<tr>
<td>σₙ</td>
<td>16.994</td>
<td>0.756</td>
<td>22.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>σₙₙ</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.310</td>
<td>0.042</td>
<td>7.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>σₙₙ</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-0.023</td>
<td>0.106</td>
</tr>
<tr>
<td>L</td>
<td>-5852</td>
<td>-2979</td>
<td></td>
<td>-6889</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lₘₙₙ</td>
<td>-5292</td>
<td>-2990</td>
<td></td>
<td>-6889</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ρ</td>
<td>0.455</td>
<td>0.723</td>
<td></td>
<td>0.343</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obs</td>
<td>9,769</td>
<td>9,769</td>
<td></td>
<td>9,769</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.11: MSL Results for synthetic data sets, where available.

7.1.4 MSL models of synthetic data sets

Correctly specified MSL models were only possible to estimate for specifications with a normal or lognormal distributed term. The results should be compared to MNL model results (table 7.8) to decide whether adding a distributed term improved the model fit. The likelihood for the MNL models (Lₘₙₙ) are added to the MSL results in table 7.11.

For the N data the estimate of the standard error of the normal distributed term σₙ, is significantly different from 0. Looking more into the estimates, it can be seen that the deviation between the ‘true’ values (1.5,0.5,1,1) is more than just a matter of scaling as ratio of estimate to ‘true’ ranges from 5.6 to 62.5.

For the LN data the model fit is improved; the improvement is significant by the likelihood ratio test. The (standard error of the) normal term (σₙₙ) is significantly different from 0, but estimated lower than the ‘true’ value (1).
CHAPTER 7. APPLICATION

Table 7.12: Significance of distributed terms, different distributions of error terms (N + LN) and different ‘true’ distribution (F through SP).

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>N</th>
<th>LN</th>
<th>TP</th>
<th>ND</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>$\mathcal{L}$</td>
<td>-5927</td>
<td>-7558</td>
<td>-4082</td>
<td>-6069</td>
<td>-9739</td>
<td>-3059</td>
</tr>
<tr>
<td>LogNormal</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>$\mathcal{L}$</td>
<td>-4205</td>
<td>-5205</td>
<td>-2979</td>
<td>-4277</td>
<td>-4277</td>
<td>-2265</td>
</tr>
<tr>
<td>$\mathcal{L}_{MNL}$</td>
<td>-4206</td>
<td>-5292</td>
<td>-2260</td>
<td>-4278</td>
<td>-6889</td>
<td>-2271</td>
</tr>
</tbody>
</table>

Note that if comparing to the variance in the data design, the variance of the distributed term is $\sigma_{LN}$ times the squared logit scale. For the ND data, the variance of the normal term is not significant – and the likelihood value is that of the MNL model. Results listed here are based on 100 draws for the random term – increasing the number of draws lowers the variance on the estimates (and may slightly change the coefficient estimate) but generally does not change whether terms are significant or not, just as changes to the likelihood value are just minor.\(^7\)

Simply for the sake of getting a grasp of whether models specified with wrong distribution for the coefficients were rejected in a MSL estimation or estimates poorly identified, models with wrong distribution were formulated. Results are presented below (table 7.12) just as whether estimates of variance are significantly different from zero or not (at 95% level). A ‘+’ indicates significance, ‘÷’ indicates non-significance. Quite surprisingly, the normal distribution is significant (or rather, the variance term is significantly different from zero) for the fixed coefficient data ($F$) as well as for all data sets with a distributed coefficient ($N$, LN, TP, SP) and the data set with a normal distributed term added (ND).\(^8\)

Furthermore, the variance of the lognormal term is significantly different from zero, for the N, LN and SP data sets. Common denominator for these sets, is that the coefficients are distributed, even skewed for LN and SP. The significance of the normal variance term for the ND set, is a bit more tricky, but may be interpreted as the additional error term seeks to explain some of the

---

\(^7\)Models based on 100, 500 and 1000 draws have been estimated, results are similar to those reported.

\(^8\)Recapturing which type of information is provided by the $t$-test. The variable, for which the $t$-test indicates a coefficient significantly different from zero, is an indication that the product of the variable and the coefficient, contributes to explain the total variation in the data. This may be caused by that the variable itself contributes to explain the variation, or, that the variable is highly correlated to another variable (that is not included), that can explain some of the variation. The latter is termed an instrument variable.
7.1 Synthetic data

Table 7.13: Significance of distributed terms by distribution (combinations of N and LN), for the data where choice is based on two attributes.

<table>
<thead>
<tr>
<th></th>
<th>F + F</th>
<th>F + LN</th>
<th>N + F</th>
<th>N + N</th>
<th>LN + LN</th>
<th>F + N</th>
<th>LN + F</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\xi$</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>$\xi_2$</td>
<td></td>
<td>+</td>
<td>$\div$</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>$L$</td>
<td>-4935</td>
<td>-4855</td>
<td>-6417</td>
<td>-13,329</td>
<td>-4852</td>
<td>-7083</td>
<td>-5104</td>
</tr>
</tbody>
</table>

variation in the residual term. Comparing the likelihood of the MSL models with the likelihood for the MNL, the picture is less dramatic, as the MNL is 'best' for the $F, LN, TP$ whereas the Lognormal MSL specification is best for $N, ND, SP$. The Normal MSL does not have the highest likelihood value for any of these data sets.

Following the path of estimating deliberately mis-specified models, the data with two attributes is pursued. All combinations of normal and lognormal distributed and fixed coefficients are estimated, including the $N + LN$ (the correct specification) and the trivial $F + F$ (the MNL formulation). Unfortunately, models including two different shapes of distributions ($N + LN$, $LN + N$) did not converge, wherefore significance is not reported here. Significance or variance terms is reported in table 7.13, where '+' ('-') indicates that the variance term is included in the model and is significant (non-significant) at a 95 % level. In addition, the likelihood value is given.

Again, most of the variance terms come out significant, only exemption is the model with two normal terms. Comparing the likelihood to the MNL estimation (the 'F + F' model), narrows the list to just two specifications, the 'F + LN' and the 'LN + LN', with close to similar likelihood value. In a practical application, the 'F + LN' could be preferred simply because of the lower number of coefficients to be estimated all though, the likelihood ratio test ($-2(L^R - L^T) = 6.9$) distributed $\chi^2$, indicate the 'LN + LN' specification is better than the 'F + LN'. Looking at the 'F + LN' and 'N + F', the lognormal distributed terms seems to be more important to specify correctly than the normal distributed term simply because of the increase in likelihood for the 'F + LN' specification, whereas only specifying the normal term correct implies a decrease in likelihood compared to the MNL.\textsuperscript{9}

If these results holds in general, there may be a problem with the way model formulations are formed/ tested, as the $t$-statistic is generally used as a key statistic. In 'real' data, the significance of a normal variance term, may be caused by some attribute(s) left unaccounted for (e.g. data was not available), where the contribution from this particular attribute to explaining the varia-

\textsuperscript{9}Further discussion on the $t$-test follows later.
tion, may be described by a normal distributed term. However, in the case of synthetic data, the formation of choice is known exactly (by the construction of the utility function). However, in principle, some other formation of data may result in the same data set, estimating close to the same coefficients, though this chance is remote.

Rather than lingering on this point, this thesis continues with assessing the distribution by the SODA method proposed in chapter 6.

### 7.1.5 Results by SODA method

Quite a few models have been estimated and to keep track of the results, models based on 50 groups (10 repetitions) are presented first in greater detail for each of the choice specifications.\(^\text{10}\) Then follows an overview of all model runs (from 50 to 500 groups, each 10 repetitions) and comparisons of the results. Charts are produced by SAS, which at times is rather rigid in the design of output. Bars of histograms are represented by midpoints and all values lower (higher) than the lowest (highest) midpoint is included in the lowest (highest) interval on the chart. Horizontal axes are attempted identical where possible, though not at all places due to large differences in mean and/or variance.

**Fixed coefficients, F1 & F3**

This model is formed for the sake of comparison, to detect the amount of variation in estimates due to the method. Results of the model where variation of the coefficient estimate is allowed on the constant terms as well as on the coefficient (F3) and the model with fixed constant terms and variation only on the coefficient \(\beta\) (F1), are shown below as figures and tables.

As this model is based on fixed coefficients and the only disturbance in the model is captured by the residual term \(\epsilon\), the distribution should be degenerate at the mean value. The spread around the mean that can be seen is due to the random sampling (group generation). The distribution of the coefficient estimate is shown in figures 7.2 (one coefficient) and 7.3 (three coefficients). The horizontal axes are identical, wherefore it is easily seen than the distributions are quite similar. The 'true' mean value (-1.0) is clearly reproduced.

The shape of the distribution resembles that of a normal, uni-modal, bell shaped, not skewed and no kurtosis. This is supported by the estimated moments, presented in table 7.14 below.

Estimation of coefficients \(\beta\), of the logit model is in fact estimation of scaled

\(^{10}\text{A low number of groups resembles having many observations in each group whereby assessments of distributions are less detailed. The advantage is lower estimation time, as this is proportional to number of groups. Repetitions of the method aids in finding patterns of distribution that are generally valid. Further discussions of the method are in section 6.}\)
7.1 Synthetic data

Figure 7.2: Distribution of $\beta$, model $F_1$

Figure 7.3: Distribution of $\beta$, model $F_3$
coefficients as the scale is set to one for identification of the model. An important issue is whether the distribution depicted of the $\beta$ is simply a matter of different scaling within the different groups of data. Had such been the case, a cross plot of coefficient estimates would show a line through the Origin for all plots (the slope equal to the ratio of the coefficients)\textsuperscript{11}. The three cross plots ($ASC_1 \times ASC_2$, $ASC_1 \times \beta$ and $ASC_2 \times \beta$) are shown below in figures 7.4 through 7.6 and do not show evidence that the distribution is caused simply by scaling.

Estimated moments for the $F1$ and $F3$ models are presented below in table 7.14, moments are calculated as measures of between-group variation (without weighting by group size). The first two moments (mean and variance) needs no further introduction. The traditional (Pearson) measure of skewness is applied here, which is given by the third moment divided by the cubic of the spread $b_1 = \hat{m}_3/(\hat{m}_2)^{3/2}$ ($m_k$ is the $k$th central moment, $\sum_{j=1}^{N}(X - \bar{X})^k/N$). $b_1 = 0$ corresponds to the normal distribution.

The measure of kurtosis employed ($g_2$), measures the departure from normality,

\textsuperscript{11}Different scaling of the coefficients in turn means different variance of the residual term. A larger scale and thereby larger variance (the variance of the error term is $\pi^2/6\mu$) will mean that even points with a larger deviation from the line the further from the Origin, will comply with scaling being the reason for the distribution of the coefficients and not differences in preferences. Data points close to a line not necessarily through the Origin, for one (few) is evidence of correlation between coefficients – not scaling.
7.1 Synthetic data

Figure 7.5: F3, plot of ASC\textsubscript{1} vs. $\beta$

Figure 7.6: F3, plot of ASC\textsubscript{2} vs. $\beta$
is $g_2 = (\hat{m}_4 / \hat{m}_2^2) - 3$, again $g_2 = 0$ corresponds to the normal distribution. Moments are further discussed in section 6.3.3; following the discussion in there, test statistics for test of moments significantly different from zero are not shown.

In the one coefficient model, $ASC_1, ASC_2$ are constrained to the value estimated for the MNL model (see table 7.8) to reflect an actual estimation where the 'true' value is unknown. In both cases the estimates of the moments of $\beta$ are quite similar and they are close to the true value. Recall, that the 'true' higher order moments for the fixed coefficients are identical 0 (as $\beta = \bar{\beta}$ at all times).

Similarly, the estimated values for the alternative specific coefficients ($F3$ model) are quite similar to the 'true' values, though they are distributed around the mean value. Their distributions are uni-modal, bell shaped and resembles the normal (which again may be explained by the normal distribution of the estimator). The distributions are shown in figures 7.7 and 7.8.

**Normal RCL, N1 & N3**

The distributions of the $\beta$ coefficients, are uni-modal, bell shaped, and not clearly skewed to either side, see figures 7.9 and 7.10. The distributions are quite similar irrespective of whether distribution was allowed for the constants or not. However, the mean of all the estimated coefficients are dislocated for both the $N3$ and $N1$ models, where the $N1$ model were based on the MNL estimates for the constants. Mean values were 15-27 % (24 %) lower compared to the true coefficients for the three (one) coefficient model\textsuperscript{12}. As the deviation was different for the different coefficients, this is not solely caused by the scale parameter. Again, it is verified that the distribution of coefficients is not due to the scale parameter $\mu$ (figures not included).

The estimated variance is quite low at 0.01 for both $N1$ and $N3$ (see table 7.15, which is at level with that of the Fixed choice (see table 7.14). Skewness and kurtosis were both higher than for the corresponding $F$ models.

\textsuperscript{12}The estimated mean of the $\beta$ coefficient is -1.02 as stated in table 7.7
7.1 Synthetic data

Figure 7.7: Distribution of ASC$_1$, model $F3$

Figure 7.8: Distribution of ASC$_2$, model $F3$
Figure 7.9: Distribution of $\beta$, model $N1$

Figure 7.10: Distribution of $\beta$, model $N3$
7.1 Synthetic data

<table>
<thead>
<tr>
<th>Coeff</th>
<th>Mean</th>
<th>Var</th>
<th>Skew</th>
<th>Kurt</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASC1</td>
<td>1.27</td>
<td>0.07</td>
<td>0.15</td>
<td>0.43</td>
</tr>
<tr>
<td>ASC2</td>
<td>0.36</td>
<td>0.07</td>
<td>-0.06</td>
<td>0.27</td>
</tr>
<tr>
<td>β</td>
<td>-0.76</td>
<td>0.01</td>
<td>-0.76</td>
<td>1.52</td>
</tr>
</tbody>
</table>

Table 7.15: Results for synthetic data where data was generated with Normal RCL specification.

Log Normal RCL, LN1 & LN3

Adding a lognormal disturbance to the coefficient allow treatment of a 'bad' ('good') characteristic of travel as such in the utility function, leading to a sign specific negative (positive) coefficient. The skewness incorporated allows some individuals to be very much more opposed than others, those being little opposes having a value close to 0, though negative.

The estimation results were quite nice. All estimates had the same sign with a distribution around the mean (the mean is -1.83). The distribution is skewed with a long tail to the left. Again, the distributions are similar for the LN3 and LN1 models. The plots of pairs of coefficients were far from a linear relation, for any pair of coefficient estimates. Hence, the depicted distribution was not caused by the scale parameter $\mu$.

Quantification of the distribution is aided by the four moments of the estimates, which are shown below in table 7.16. The mean is estimated 15-16\% below the true mean value, whereas the constants are only estimated 2-4 \% below. The reason for the smaller $\beta$ estimates is that records with a very high (absolute) true value are grouped with records with lower true values, due to the random grouping procedure. The estimates will be somewhere in between as the estimation implicitly assumes a uniform value for the group. Hence, the extreme values of $\beta$ will only be realised by estimation if a group consists only of extreme values! That is not to say that all values will be equal nor they will be evenly distributed around the mean.

Twin Peaks RCL, TP1 & TP3

In a segmented population the distribution of preferences is degenerated, i.e. each segment is described by the same coefficient, given that the population within each segment possesses identical preferences. If on the other hand, individuals’ preferences are different, each segment may be described by a distribution, say uni-modal. The whole data set may then be described by the collection of these distributions, a $t$-modal distribution, where $t$ is the number of segments.

The lowest estimated coefficient value is -2.47 respective -2.56 for the LN1 respective LN3 model.
Figure 7.11: Distribution of $\beta$, model LN1

Figure 7.12: Distribution of $\beta$, model LN3
7.1 Synthetic data

<table>
<thead>
<tr>
<th>Coeff</th>
<th>3 Coeff</th>
<th>Mean</th>
<th>Var</th>
<th>Skew</th>
<th>Kurt</th>
<th>1 Coeff</th>
<th>Mean</th>
<th>Var</th>
<th>Skew</th>
<th>Kurt</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASC1</td>
<td>1.43</td>
<td>0.14</td>
<td>0.11</td>
<td>0.20</td>
<td></td>
<td>1.46</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>ASC2</td>
<td>0.49</td>
<td>0.14</td>
<td>0.11</td>
<td>0.47</td>
<td></td>
<td>0.46</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>β</td>
<td>-1.54</td>
<td>0.05</td>
<td>-0.86</td>
<td>1.46</td>
<td></td>
<td>-1.53</td>
<td>0.04</td>
<td>-0.79</td>
<td>1.67</td>
<td></td>
</tr>
</tbody>
</table>

Table 7.16: Results for synthetic data where data was generated with Log-Normal RCL specification.

<table>
<thead>
<tr>
<th>Coeff</th>
<th>3 Coeff</th>
<th>Mean</th>
<th>Var</th>
<th>Skew</th>
<th>Kurt</th>
<th>1 Coeff</th>
<th>Mean</th>
<th>Var</th>
<th>Skew</th>
<th>Kurt</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASC1</td>
<td>1.49</td>
<td>0.10</td>
<td>0.31</td>
<td>0.29</td>
<td></td>
<td>1.46</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>ASC2</td>
<td>0.49</td>
<td>0.10</td>
<td>-0.04</td>
<td>0.21</td>
<td></td>
<td>0.46</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>β</td>
<td>-1.00</td>
<td>0.01</td>
<td>-0.38</td>
<td>0.47</td>
<td></td>
<td>-0.98</td>
<td>0.01</td>
<td>-0.45</td>
<td>0.84</td>
<td></td>
</tr>
</tbody>
</table>

Table 7.17: Results for synthetic data where data is generated with Twin Peaks RCL specification.

Taking out random groups of data, often contains records from both these segments, whereby neither of the true values are likely to be estimated. Similarly, it is likely that the vast majority of records in some groups will belong to either of the segments whereby an estimate closer to this segment’s ‘true’ value is likely. As may be seen from the graphs the twin peaked distribution is not recovered: Looking just at the table, it can be seen that it fits the mean value nicely, \( \beta \) is less than 2 % lower than the true value and the constants fit within 2 %. The distributions and calculated moment are fairly similar for the TP3 and TP1 models.

**Normal Disturbance added, ND1 & ND3**

In another specification seen in the literature (Logit Kernel with a probit) both a Gumbel distributed and a Normal distributed error terms occur. Such a data set was constructed, though it was not expected that this method could detect this, as the general formulation from the above methods were used.

For data with more than one record per individual, this model may be tested using the SODA method, with utility functions specified as

\[
U_n = ASC_i + \beta X + \eta_n + \epsilon_n, \forall n < N, ASC_i = 0, \text{ for } i = 3 \quad (7.9)
\]

with \( ASC_i, \beta \) fixed over model runs and \( \eta \) estimated for each group. This will give individual specific parameters, that may by used for obtaining the distribution.
Figure 7.13: Distribution of $\beta$, model $TP1$

Figure 7.14: Distribution of $\beta$, model $TP3$
7.1 Synthetic data

<table>
<thead>
<tr>
<th>Coeff</th>
<th>3 Coef</th>
<th>1 Coef</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Var</td>
</tr>
<tr>
<td>ASC1</td>
<td>0.85</td>
<td>0.05</td>
</tr>
<tr>
<td>ASC2</td>
<td>0.34</td>
<td>0.05</td>
</tr>
<tr>
<td>β</td>
<td>-0.51</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 7.18: Results for synthetic data where data is generated with Normal disturbance added.

Model results are listed below in table 7.18. The ND3 and ND1 models does not recover the model coefficients and this is not (solely) a matter of scaling (estimates are 32 - 49 % for ND3, and 44 % lower for ND1). The variance is quite small while skewness and kurtosis are small, in fact all three moments are smaller than those of $F_3$ and $F_1$ models, respective. A likely explanation is that the estimation assumption of records from a normal population is fulfilled.

'Stated Preference', $SP_1$ & $SP_3$

By construction of the data – resemblance of a Stated Preference data set with repetitive choice situations for each respondent, repetition by alternative group formation is not possible (groups are formed with exactly one respondent in each group, each respondent only belonging to one group). Models based on 50 groups have not been estimated, as this resembles having 200 records for each respondent. In an actual data set, having 200 records per respondent is unrealistic, wherefore the model based on 50 groups has not been pursued any further. Estimation results (distribution) for the 500 groups case (20 records for each respondent) are given below in figures 7.17 and 7.18. Do note the changed axis values on the horizontal axes and that these charts are only based on one 'repetition'.

The distributions for the $SP$ models are far wider than for the other distributions.

N + LN RCL, $NLN_2$ & $NLN_4$

The final synthetic data set was based on two attributes; both of these were generated by a normal distribution (different mean, variance and seed). The first coefficient was assumed normal distributed, while the second was assumed lognormal distributed (different mean, variance and seed). The purpose of using two different distributions is to assess whether the method can distinguish between the two distributions, identify two different shapes and identify which has the higher variance and mean value, correctly.
Figure 7.15: Distribution of $\beta$, model $ND1$

Figure 7.16: Distribution of $\beta$, model $ND3$
7.1 Synthetic data

Figure 7.17: Distribution of $\beta$, model $SP1$ (based on 500 groups)

Figure 7.18: Distribution of $\beta$, model $SP3$ (based on 500 groups)
Using two different distributions to describe individuals actions corresponds to admitting that individuals are different, but their perception of one attribute is more different that their perception of the other attribute (different variance). A symmetrical distribution (e.g. the normal) is employed when individuals are evenly spread on either side of the mean (most are close to the mean), whereas a skewed distribution (e.g. the lognormal) is used when there is more individuals on one side of the mean and/or these are 'further away' from the mean value.

The charts of distribution of $\beta$ and $\beta_2$ (figure 7.19 through 7.22) are very similar. The difference between the models with/without distribution on the ASC’s is limited compared to the one-attribute models. Furthermore, the differences in the shape of distributions, as well as the difference in the variance levels ($var(\beta_2) > var(\beta)$) are both recovered.

The estimation result (table 7.19), which were quite similar for the specifications where two respective four coefficients, were estimated, in terms of estimates, variance and higher order moments. This is an indications that fixing the constants has less impact when there is more than one attribute. Comparing to the 'true' values ($ASC_1, ASC_2, \beta, \beta_2) = (1.5, 0.5, -1.02, -3.39)$, these estimates are far from the 'true' values in absolute terms, as well as the ratios. This confirms that, when coefficients are distributed, a fixed coefficient specification cannot correctly estimate the coefficients.

### 7.1.6 50-500 groups, 1 and 3 coefficients

The preceding section described results for each data set, this section will focus on comparison of results as well as summarise a number of model runs. The models have been run based on 50, 100, 250 and 500 groups, though the $SP$ is just reported for the 500 group case (recall that the data is constructed to mimic 500 individuals’ behaviour – hence, smaller number of groups or regrouping of observations is not an option). Three tables below (7.20 through 7.22) lists mean, variance and median for the model runs, for models where 3 (upper

<table>
<thead>
<tr>
<th>Coeff</th>
<th>4 Coeff</th>
<th>2 Coeff</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Var</td>
</tr>
<tr>
<td>$ASC_1$</td>
<td>1.01</td>
<td>0.0975</td>
</tr>
<tr>
<td>$ASC_2$</td>
<td>1.25</td>
<td>0.1210</td>
</tr>
<tr>
<td>$\beta$</td>
<td>-0.41</td>
<td>0.0029</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-1.09</td>
<td>0.0215</td>
</tr>
</tbody>
</table>

Table 7.19: Results for synthetic data where choice is generated from a two variable RCL (Normal and Log-Normal) specification.
7.1 Synthetic data

Figure 7.19: Distribution of $\beta$, model $NLN2$

Figure 7.20: Distribution of $\beta$, model $NLN4$
Figure 7.21: Distribution of $\beta_2$, model NLN2

Figure 7.22: Distribution of $\beta_2$, model NLN4
### 7.1 Synthetic data

<table>
<thead>
<tr>
<th>Groups</th>
<th>N1</th>
<th>LN1</th>
<th>TP1</th>
<th>ND1</th>
<th>F1</th>
<th>SP1</th>
</tr>
</thead>
<tbody>
<tr>
<td>True value</td>
<td>-1.02</td>
<td>-2.14</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-2.13</td>
</tr>
<tr>
<td>50</td>
<td>-0.78</td>
<td>-1.53</td>
<td>-0.98</td>
<td>-0.56</td>
<td>-1.00</td>
<td>-</td>
</tr>
<tr>
<td>100</td>
<td>-0.79</td>
<td>-2.11</td>
<td>-1.00</td>
<td>-0.57</td>
<td>-1.02</td>
<td>-</td>
</tr>
<tr>
<td>250</td>
<td>-0.84</td>
<td>-2.46</td>
<td>-1.05</td>
<td>-0.58</td>
<td>-1.07</td>
<td>-</td>
</tr>
<tr>
<td>500</td>
<td>-0.95</td>
<td>-2.74</td>
<td>-1.20</td>
<td>-0.61</td>
<td>-1.23</td>
<td>-2.52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Groups</th>
<th>N3</th>
<th>LN3</th>
<th>TP3</th>
<th>ND3</th>
<th>F3</th>
<th>SP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>True value</td>
<td>-1.02</td>
<td>-2.14</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-1.00</td>
<td>-2.13</td>
</tr>
<tr>
<td>50</td>
<td>-0.76</td>
<td>-1.55</td>
<td>-1.00</td>
<td>-0.51</td>
<td>-1.02</td>
<td>-</td>
</tr>
<tr>
<td>500</td>
<td>-2.84</td>
<td>-13.18</td>
<td>-4.64</td>
<td>-0.77</td>
<td>-4.46</td>
<td>-14.79</td>
</tr>
</tbody>
</table>

Table 7.20: Mean of estimated coefficients, all data sets.

<table>
<thead>
<tr>
<th>Groups</th>
<th>N1</th>
<th>LN1</th>
<th>TP1</th>
<th>ND1</th>
<th>F1</th>
<th>SP1</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>-0.77</td>
<td>-1.52</td>
<td>-0.98</td>
<td>-0.56</td>
<td>-1.00</td>
<td>-</td>
</tr>
<tr>
<td>100</td>
<td>-0.77</td>
<td>-2.00</td>
<td>-0.98</td>
<td>-0.57</td>
<td>-0.99</td>
<td>-</td>
</tr>
<tr>
<td>250</td>
<td>-0.80</td>
<td>-2.07</td>
<td>-0.99</td>
<td>-0.57</td>
<td>-1.02</td>
<td>-</td>
</tr>
<tr>
<td>500</td>
<td>-0.80</td>
<td>-2.09</td>
<td>-1.00</td>
<td>-0.55</td>
<td>-1.02</td>
<td>-1.94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Groups</th>
<th>N3</th>
<th>LN3</th>
<th>TP3</th>
<th>ND3</th>
<th>F3</th>
<th>SP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>-0.75</td>
<td>-1.51</td>
<td>-0.99</td>
<td>-0.51</td>
<td>-1.02</td>
<td>-</td>
</tr>
<tr>
<td>500</td>
<td>-0.90</td>
<td>-3.41</td>
<td>-1.21</td>
<td>-0.57</td>
<td>-1.25</td>
<td>-3.71</td>
</tr>
</tbody>
</table>

Table 7.21: Median of estimated coefficients, all data sets.

The general tendency is that mean of the estimates increases as the number of groups increases. This conforms to the intuition, as more extreme coefficient estimates are more likely to occur as group size diminishes. To identify whether the density mass of distribution has moved or it is just some extreme estimates that has become 'more extreme' and thereby affects the mean value, the median (50 % of the observations are on either side of the median) is used. The median shown in table 7.21, is by and large stable for the 1-coefficient models, though the LN1, 50 groups stands out. This supports that the mass density is unaltered for values above the median (closest to zero) and the left hand tail has been 'longer'. The median for the TP and F model are at the correct level and LN almost at the correct level (compared to the 'true' coefficient, in table 7.7). The N model is 20 % below the 'true' value while the ND is just half of the 'true' value.
The case is somewhat different for the 3-coefficient model, as the estimated medians shift to numerical higher values by a factor around 20% for the $N$, $TP$ and $F$ models. The $ND$ model 'only' shift up by 13% while the $LN$ model more than doubles (125%).

Unanimous growth in the variance seems to be the case, corresponding to that more estimates far from the mean, are realised as the groups size decreases (number of groups increases). The picture is somewhat more explosive for the 3-coefficient models. Again, in the 3-coefficient model the two ASC’s may cover up some of the variation in the $\beta$ or even some of the variation in the $\epsilon$. The latter would explain the realisation of the extreme values of the $\beta$ estimates, whereas the former is more likely to explain the change of sign in the ASC’s. It should be mentioned that the estimate of variance presented here (between-group variance) is not comparable to the variance of the estimated coefficients in the MNL model in table 7.8 (variance of the $\beta$ estimates).

Returning for a moment to the logit scale, which all coefficient estimates are multiplied by (due to identifiability of the model) and hence, may play a role in these results as well. The logit scale ($\mu$) is related to the variance of the error term $\epsilon$, by $\mu = \sqrt{(\pi^2)/6 \, \text{var}}$, that is, if the variance increases the scale decreases. Looking at the mean of the estimates for the $N$ and $ND$ models, in this light, a possible explanation for this is that the variation of the error term has increased (decreasing scale by increasing number of groups); in particular for the $ND$ model. The $ND$ model has a normal error added which the utility used in the modelling is not designed to capture, hence the error term seek to overcome this resulting in a higher variance of the $\epsilon$. A scale of 0.55 (the ratio of the estimate and the 'true') corresponds to a variance of 5.4; a scale of 1 corresponds to a variance of 1.65.

The variance of the error term reflects the difference between the observations the model is based on. Estimating different models for groups of the data, where observations are assigned to groups at random, may increase or decrease
the variance. If the observations are similar (in 'true' coefficient values) the variance will go down (the scale will go up) – reflected by higher absolute values of the coefficients (everything else being equal). Groups where observations represents the range of the 'true' coefficients will have a high variance – hence, lower absolute values of the coefficients. To detect whether the revealed variation in coefficient values is simply due to scaling scatter-plots of coefficients are produced\(^\text{16}\).

In figures 7.23 and 7.24 below, scatter plots of $\beta$ and $ASC_1$, (horizontal, vertical). All models are presented in the same figure to clearly show the difference between models. The figure of models based on 50 groups (figure 7.23) does not show any linear relation between the estimates of $\beta$ and $ASC_1$ (the plot for $\beta \times ASC_2$ is similar, not included here). Hence, there is no evidence that the variation in estimated coefficients is due to scaling. The shift in location of the mass density, between the models, implies different relative tradeoffs between models coefficients.

The picture is somewhat different for models based on 500 groups (figure 7.24) where the points are more spread over an elongated area. As the values gets further away from the 'true' (corresponding to a scale of 1), they are generally close to the line through the Origin with slope equal to the ratio of the 'true' values. The reference line shown has slope 0.67 roughly equal to the trade off for the $TP$, $ND$, $F$ and $N$ models. The corresponding line for $LN$ and $SP$ models is twice as steep.

Looking at the ratios of coefficients $\beta/ASC_1$ the picture changes. The ratio of the estimated coefficients is calculated as the ratio $\beta/ASC_1$ for each group, before calculating the average over groups. Hereby the effect of the hidden logit scale is eliminated and furthermore, no restrictions on the distributions for the each of the coefficients $\beta, ASC_1$ is required. The ratio of the estimated coefficients is presented next to 'true' ratio’s in table 7.23. For a symmetric distribution the mean and the median are equal, whereas for positively skewed distributions the median is smaller than the mean, as the mean is largely affected by the tail.

The median and the mean values are fairly close for the 50 group case, while quite different for the 500 group case. As the mean generally increases, the median generally decreases (increases slightly for $LN$ model). This in turn, suggest that the mass density (of the ratio) is not affected be the increase in number of groups, but some extreme values causes the increase in the mean value. Extensive scaling between data groups, may account for most of the variation depicted.

\(^{16}\text{Plots of } \beta \text{ against } ASC \text{ are just produced for the 3-coefficients model; plots for the 1-coefficient model would not reveal scaling as the } \beta \text{ is estimated conditional on the (fixed) value of the } ASCs.\)
Figure 7.23: Scatter plot of $\beta$ (horizontal axis) and $ASC_1$ (vertical axis), $F$, $LN$, $F$, $TP$ and $ND$ models for 50 groups.

Figure 7.24: Scatter plot of $\beta$ (horizontal axis) and $ASC_1$ (vertical axis), $F$, $LN$, $F$, $TP$ and $ND$ models for 500 groups.
7.1 Synthetic data

<table>
<thead>
<tr>
<th></th>
<th>'True' values</th>
<th>Estimates, 50 groups</th>
<th>Estimates, 500 groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E \frac{\beta_{ASC1}}{Med \frac{\beta_{ASC1}}{}}$</td>
<td>$E \frac{\hat{\beta}<em>{ASC1}}{Med \frac{\hat{\beta}</em>{ASC1}}{}}$</td>
<td>$E \frac{\hat{\beta}<em>{ASC1}}{Med \frac{\hat{\beta}</em>{ASC1}}{}}$</td>
</tr>
<tr>
<td>$N$</td>
<td>-0.68</td>
<td>-0.62</td>
<td>-2.08</td>
</tr>
<tr>
<td>$LN$</td>
<td>-1.43</td>
<td>-1.16</td>
<td>-1.69</td>
</tr>
<tr>
<td>$TP$</td>
<td>-0.67</td>
<td>-0.70</td>
<td>-1.81</td>
</tr>
<tr>
<td>$ND$</td>
<td>-0.67</td>
<td>-0.66</td>
<td>-0.75</td>
</tr>
<tr>
<td>$F$</td>
<td>-0.67</td>
<td>-0.72</td>
<td>-0.25</td>
</tr>
<tr>
<td>$SP$</td>
<td>-1.42</td>
<td>-1.33</td>
<td>-0.64</td>
</tr>
</tbody>
</table>

Table 7.23: Ratio of coefficients, 'true' and estimates, all data sets, 50 and 500 groups.

A final look at the distributions where a removal of the scale effect has been attempted, by plotting ratios of $\beta$ and the ASCs. The figures 7.25 through 7.29 shows the distribution with a fitted Normal (dotted blue) and Lognormal (solid red). Values below 0 and above 10 have been truncated in order to show the fitted distribution for the normal and log-normal distributions. For all models the lognormal seem to have a better fit than the normal, even for those data sets where the 'true' distribution (of the coefficients) is known to be symmetrical.

7.1.7 Comparisons of synthetic data

The above seven data sets are examples of how actual data may be constituted. The analysed data sets covered different shapes of distributions, different levels of variance, a data example set with repetitive choice situations for each 'individual' and finally an example of data set where two attributes (distributed coefficients with different shape) were included. Many more formulations could have been tested, just not within the time frame for this study.

An important lesson learned from this application is that testing by means of synthetic data, is not 'just done'. The amount of time spent is significant, as increased insight enables the researcher to perform increasing difficulty of tests in the search for 'unbiased conclusions' – that is, conclusions where in principle, any other explanation is ruled out. On the other hand, if or when the test performs as expected, the joy is plentiful.

The constructed data was analysed prior to the test with the SODA method. This included changes in 'individual' choice, between different shapes of distributions, estimation by a traditional fixed coefficient specification – a MNL model, and finally, a MSL estimation where possible (limitation of shape of distributions). For the sake of the experiment, MSL models were also estimated assuming an incorrect shape of the distribution.

All MNL models had fairly high levels of significance for the two constants.
Figure 7.25: Distribution of $\beta/ASC_1$ for $N$, 500 groups

Figure 7.26: Distribution of $\beta/ASC_1$ for $LN$, 500 groups
7.1 Synthetic data

Figure 7.27: Distribution of $\beta/ASC_1$ for $TP$, 500 groups

Figure 7.28: Distribution of $\beta/ASC_1$ for $ND$, 500 groups
CHAPTER 7. APPLICATION

Figure 7.29: Distribution of $\beta/ASC_1$ for $F$, 500 groups

Figure 7.30: Distribution of $\beta/ASC_1$ for $SP$, 500 groups
7.1 Synthetic data

and the coefficient. Not surprisingly, the model on data based on fixed coefficients ($F$) reproduced the coefficients, but also did the model on data based on TwinPeak distributed coefficients ($TP$) and the two data sets with lognormal distributed coefficients added ($LN$ and $SP$). The model based on choice generated from normal distributed coefficients ($N$) and a normal disturbance added ($ND$), roughly reproduced the ratios of coefficients. Results from the two-attribute model ($NLN$) did not reproduce 'true' values nor ratios of coefficients, even the relative size of constants was incorrect.

Estimation of models by the MSL method was limited to the normal and lognormal distributions (software restriction). The variance term for the distributed coefficient, was significantly different from zero for the $N$ and $LN$, but insignificant for the $ND$ model. Differences between the 'true' coefficients and the estimates were a factor of 5 to 63; or stated differently, the results were implausible. For the $LN$ model, the constants were correct, yet, mean and standard error were estimated at less than half of the 'true' coefficient value. However, these results are from one-attribute models, where interaction with the residual term may be more outspoken than for models where choice is described by several attributes. Unfortunately, for the two-attribute model ($NLN$), the correct specification has not yet been possible to estimate by MSL (optimisation problems within the program), wherefore comparisons of estimates and 'true' values are unsettled.

To get a hunch of how reluctant the MSL estimation method was to indicate significance of distributed terms, models of wrong assumed distribution were estimated. Generally, the normal distribution of coefficients was significant, irrespective of how the choice data was formed, only exemption was the two-attribute model with two normal distributed terms. The case was different for the lognormal distribution, as this was only significant for data sets where uni-modal distributed coefficients were used in the data generation. The lognormal distribution was significant for all tested specifications of the two-attribute model, that is two lognormal ($LN + LN$) or one lognormal combined with a fixed coefficient ($F + LN, LN + F$).

Looking closer at the interpretation of the $t$-test, it reveals that significance of some term, indicates that, that particular terms can explain some of the total variation in the data. As the added term is an estimator of the variance (for some distributed term), the $t$-test does gives us an indication whether this variance term can explain some of the variation in data (is significantly different from zero), in the sense that the model is improved by it's presence. But, it does not give any information on whether the a priori assumed shape of distribution for that term, is the 'correct' shape or not! In essence, the $t$-test cannot be used as a key statistic for testing whether the 'correct' shape of distribution has been

---

17 The correct absolute values were recovered if the logit scale was 0.75 respective 0.55 for $N$ respective $ND$ data. Scaling less than one indicate a higher than unit variance of the residual term, which indicate that the variance on the coefficient is captured by the residual term.
The empirical assessment of shape of distribution was performed for all generated data sets, with and without including the constants; and by data partitioned into different number of groups (50-500). In general, increasing the number of groups tended to skew the distribution, although with unaltered place of mode (and median). In this case, 250 groups (40 records per group) gave the best tradeoff of accuracy and not to skewed by the number of groups.

As a whole, the charts of distributions, those included here (50 and 500 groups) and those assigned to appendix C (100 and 250 groups), were bell shaped and more importantly, they did retrieved the correct sign (sign specific). Distributions were based on estimates from all groups (it was possible to obtain estimates for all groups). As the estimates are obtained from a model where scaling was included (all coefficients are scaled by the logit scale), a necessary check was to determine whether the apparent distribution of the coefficients was caused by this scale. Scatter plots of estimates did not reveal such a relation, wherefore it was concluded that the revealed distributions were not due to scaling.

The reference data set with fixed coefficients $F$, revealed some bias in results caused by the method. Had the method been unbiased, plots of $F$ should be a spike – the deviation from the mean is due to sampling (group formation). This bias should be 'subtracted' from the distributions of the other data sets.

Two data sets were constructed based on uni-modal distributed coefficients, the $N$ and $LN$. The $LN$ chart show a skewed distribution, whereas the $N$ is close to symmetrical. The distribution is wider for the $LN$ than for the $N$, which is compatible with the differences in the 'true' variation. The distribution of the $N$ is slightly slimmer than $F$, that is, it is more dense around the mode, while the distribution for $LN$ is wider than for $F$. For $LN$, there is a slight shift in location of the distribution, when distribution is allowed for the constants (from $LN1$ to $LN3$).

The $SP$ did also include lognormal distributed coefficients, but due to the nature of the data (repetition of choice by 'individuals'), model repetitions and plots based on 50 groups, were not an option. Charts of distribution based on 500 groups, is wider than for the $LN$, because of the number of groups. The distributions are clearly skewed. The spike at the lowest value covers over an extremely long tail, which for $SP3$ runs up to -80! The density evenly spread between -10 and -80 (1/2 to 1 % per interval of 0.5). The tail is longer than that for $LN$ - again, this is due to the number of groups.

The $TP$ was constructed as an approximation of a segmented data set, with distributed coefficients. It was not expected this would be revealed by this test, as groups are formed at random. In the charts this was depicted as a distribution slightly wider than $F$, that is wider 'shoulders' (also referred to as kurtosis).
7.1 Synthetic data

These wider shoulders, indicate a variation in 'true' values is present in data, but to which extent (segmented population, uniform distributed preferences) it does not provide information on.

The ND had fixed coefficients, but also a distributed term was added. The method searched for a distributed coefficient, and retrieved a distribution for ND slimmer than that for F. There is a shift in location between ND1 and ND3, just as the mode is dislocated compared to F.

For the NLN data the distribution for two coefficients (and two models NLN2 and NLN4) are shown. The 'true' distribution for $\beta$ is normal, the chart is close to symmetrical (both NLN2 and NLN4). It is slimmer than the distribution for N. The distribution for NLN2 is flat with 'heavy' shoulders. Furthermore, it is slightly skewed - but, the distribution is slimmer than for LN, though the 'true' variance of LN is less than for $\beta_2$. The distribution is much wider than for F. However, the correct comparison should perhaps, be to a two-attribute (fixed + fixed) model, as in the one attribute model there is a potential, that variation has been pulled over from the residual, to an extent that is not happening for multi-attribute data.

The setting for the method (number of groups and repetitions) has been shown to have some impact on the results, as estimates are affected by the number of groups via differences in the within group variance and, hence the logit scale for that particular group’s estimates. This effect was visible through the estimates as well as by use of scatter plots of coefficient estimates. Taking out the effect of the scale, that is, looking at ratios seemed to limit the problem.

An alternative specification, where all group estimates of the coefficients were estimated jointly was pursued, though without any luck. All groups model were specified in the same utility function, such that consistent estimates of the coefficients (all coefficients multiplied by the same logit scale) were obtained. Hereby, any doubt that the variation in the estimated coefficients was due to scaling by the logit scale, could be left out. This failed as the specification was not estimable due to (very) high correlation between each of the group estimates ($\beta_g$). Whether this is a program setting in ALOGIT and whether such a model specification is possible in other software has not been investigated any further.

7.1.8 Conclusion

This test on synthetic data showed several important issues. Starting with the data formation; as data was formed the impact of the different distributions of distributed terms became evident. 'Individual' choice, did change by change of shape for distributed terms! This could be seen at 'individual' level as on aggregate level.

Estimation of the traditional MNL model, revealed what may be used as an
indicator (but not a guarantor) that additional distributed terms are needed in a model. Depending on the shape of the distribution, the size of the logit scale was higher than one, resulting in absolute coefficient estimates lower than the 'true'.

The Mixed Logit model estimation (by MSL) revealed some shortfalls of the use of the $t$-statistic. Generally, for all the estimated models, irrespective of whether these were correct or misspecified models, all estimates, that is, coefficient and variance terms, turned out significantly different from zero. Hitherto, this has been interpreted as 'significance of that particular distribution' – or, less strict speaking, that 'the correct distribution had been recovered'. This test clearly recovered that such is not the case. Merely, that significance of a variance term indicates that that particular terms can explain some of the total variation in data (which is also the correct interpretation of the $t$-statistic).

Unfortunately, the author is not aware of a statistical test that can be used in combination with Mixed Logit estimation, for testing significance of shape of distribution of distributed terms. A way to work around this problem is to specify the distributed term rigoursly as a combination of differently distributed terms, such that the $t$-test can be applied to test whether each of these terms contribute to explain to total variation. Following the notation i eq. (2.4) where $U_{ji} = V_{ji} + \eta_j + \epsilon_j$, let $\eta_j = \eta_1^j + \eta_2^j + \ldots + \eta_Q^j$, where each of the $\eta_q^j$ are formed as a coefficient multiplied by a standard distributed (mean 0) random variable, each following different shapes of distributions. The resulting distribution (removing terms with coefficients not significantly different from 0) is a parametric distribution. However, this will cause the number of coefficients in the model to increase dramatically if all possible shapes of distributions must be considered. Further, if the estimated model should later be used for forecasting, this parametric distribution must be forecast to the forecast time. An example where $Q = 2$ is found in Nunes et al. (2001), even when the mix is from this low number of distributions, the resulting distribution seems complicated.

Application of the SODA method, revealed that only having one attribute to explain choice, may cause some distortion of the variation in the estimation phase. Recovered distribution tended to be skewed, which was likely to be caused by the residual distribution (the Gumbel distribution is skewed). For the model where choice was based on two attributes, the result was by far clearer. The correct shapes (one normal and one lognormal distribution) were recovered and further, the highest level of variance was recovered for the correct term. The logit scale factor was turned down as the explanation for the depicted variance, by plots of coefficient estimates.

At this point it seems plausible that the SODA method works. Therefore, the method will be applied to 'real' data in the following section.
7.2 The Copenhagen-Ringsted case

In 1998/99 a model was developed to evaluate an infrastructure project in the Eastern part of Denmark between Copenhagen and Ringsted. However, it was designed to become a general model for future project evaluations in Denmark. The Copenhagen-Ringsted model (CRM) was designed to describe issues (in addition to what is normally covered by most 4-step models) as capacity and punctuality of public transport services as well as congestion on roads, evaluation of impacts, public transport supply (frequencies) and detailed time-schedules (real-time based), changes in route choice and demand effects due to changes in various time components and finally, differences in individuals preferences within and between segments. For more description of the CRM see (Nielsen et al., 2001).

7.2.1 Data description

The data applied for the model was partially re-used data and data collected specifically to shed light on specific issues not covered in previously collected data. The time span for data collection was 1992-98, which was covered by 3 revealed preference data sets. Furthermore 7 stated preference surveys were included – these were constituted by 29 SP experiments. A total of 16 mode alternatives were available in the CRM data, though each data set included fewer alternatives.  

The original model was estimated in several steps. At first a nested logit was applied due to the number of data sources, then a nest splitting between private and public transport modes. This adds an extra level of scaling to the model, compared to the synthetic data in the previous section. Error components (random coefficients) were added in the utility functions to facilitate modelling of individual’s different behaviour. Further distributed terms were added in the assignment of traffic though; this thesis does not consider these. A number of variants of random coefficients were tested by means of maximum simulated likelihood (Normal and Lognormal distributions applied to different terms in the utility function).

This test could have been approached in a number of different ways - a separate test for each data source, a test for either SP or RP data or a test with all data included. At that point the number and size of the data sources had to be kept in mind. Different tests were been set up, though this thesis will focus on a rather simple model described by 4 coefficients (access/egress time, cost, public transport in-vehicle time and car in-vehicle time). The results in this section were obtained solely from SP data, due to problems with the RP

18Train was sub-divided by level of service (InterCity, Regional trains, local S-trains in Copenhagen. Alternatives were treated differently in RP and SP data collections to see whether a difference in the level of information available to the individual had any effect on the preferences revealed (coefficients estimates).
data. The data applied consisted of 7 SP data sets including 26 experiments - a total of 14,708 observations after outlier removal. The estimation of the model coefficients was undertaken in ALOGIT, which made it an option to reuse a significant amount of code (data validation and model specification).

The model set-up was kept constant for all models; i.e. the same utility function was applied for all groups - though $t$-statistics for some coefficients in some model run may have had suggested removal of one (some) variable(s). All alternative specific constants were fixed to the value estimated from the NL estimation (they became $\pm\infty$ in estimation). All scales motivated by the number of data sources were fixed to the value estimated in the non-grouped case (the value estimated in the standard NL model).

The following section will provide results from the analysis as well as more practical comments on implementation/forecasting.

7.2.2 Test 1, 100 groups

This test contained the four attributes cost [Cost], access/egress time [AEtime], car travel time [Cartime] and public travel time [PTtime]. The data was divided into 100 groups by assigning random group numbers (draws from a discrete $U(1,100))$ to each record. Multiple records were available for each individual, though not enough to estimate separate models for each individuals, as the data was Stated Preference data. To work around problems of repeated measurement, records concerning the same individual were assigned group numbers as record concerning different individuals. The procedure was repeated 30 times; charts and estimated moments are based on all 30 repetitions, as a contrary to the test on synthetic data (those were based on 10 repetitions).

Charts of the distribution of the estimated coefficients are given in figures 7.31 through 7.34. Where possible, the fitted normal and lognormal (only for sign specific distributions) are shown as blue dotted respective red line. Note there are differences between the axes. The distributions of the coefficients (cost, PTtime, Cartime) are skewed as well as have heavier tails - though, the non-normality of the coefficient for AEtime is more questionable. This is perhaps more clearly illustrated by the charts below. The X-axis gives the value of the coefficient while Y-axis gives frequency (percentage).

The distribution of the coefficient for access/egress time is shown in figure 7.31. This is a bell-shaped distribution slightly skewed to the right (towards higher values). The distribution is sign specific (positive), with a low variance compared to the mean (the Coefficient of Variance CoV, is shown in table 7.24). Furthermore, the lognormal distribution (red line) seems to fit the data better than the normal distribution (dotted blue line) in the chart.

The coefficient for cost shown in figure 7.32, is bell-shaped and symmetrical.
7.2 The Copenhagen-Ringsted case

Figure 7.31: Access/ egress time

Figure 7.32: Cost
The distribution is not sign specific (approx. 3/4 are positive the rest negative), wherefore only the normal distribution is fitted. The distribution seems to be nicely described by a symmetrical distribution, and in particular the normal distribution.

Total public transport time, shown in figure 7.33, is bell-shaped and symmetrical. It appears to be described by a close to symmetrical distribution, the normal has a very nice fit of the data. The distribution for Cartime, is shown in figure 7.34. This distribution has without doubt the largest variation in the estimates (in absolute terms). The distribution is bell shaped and clearly skewed to the right. The lognormal distribution describes the variation very well –and better than the normal distribution.

To support the above charts of distributions, moments are estimated; mean, variance, skewness and kurtosis of the four coefficients, are shown in table 7.24. Note, that the scale for mean is \((10^{-1})\) while \((10^{-3})\) for the variance. The coefficient of variation (CoV) is defined as the ratio of the standard deviation to the mean. Recall, that for symmetrical distributions approximately \(2/3\) (95%) of the observations falls within \(\pm\) one (two) standard error of the mean.

Mean of estimated coefficients is highest for the car time (1.72) followed by PT time (0.69) with AE time valued about half (0.36). The median is lower for the skewed distributions though similar to the mean for the symmetrical distributions. Stated differently, the (dis)-preference for public transport time is the double of the (dis)-preference for the time it takes to get to the bus/train stop. Similarly, the travel time by car is valued at more than double of the time spent in a public transport mode.

The variance is the highest for car time, followed by cost (highest CoV); access/egress time and PT time has the lowest variance. High variation of cost (relative to the mean) covers over that some estimates are negative. This in turn causes estimates of the value of time to have both signs! The estimated

\[ \text{Table 7.24: Estimated moments, the first four} \]

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Med (10(^{-1}))</th>
<th>Variance (10(^{-3}))</th>
<th>CoV</th>
<th>Skew</th>
<th>Kurt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>0.15</td>
<td>0.14</td>
<td>0.83</td>
<td>1.97</td>
<td>-0.80</td>
<td>11.37</td>
</tr>
<tr>
<td>Access/ Egress time</td>
<td>0.36</td>
<td>0.32</td>
<td>0.21</td>
<td>0.41</td>
<td>5.80</td>
<td>99.50</td>
</tr>
<tr>
<td>Car travel time</td>
<td>1.72</td>
<td>1.67</td>
<td>2.40</td>
<td>0.28</td>
<td>0.67</td>
<td>0.72</td>
</tr>
<tr>
<td>Public transport time</td>
<td>0.69</td>
<td>0.68</td>
<td>0.32</td>
<td>0.26</td>
<td>0.21</td>
<td>0.23</td>
</tr>
</tbody>
</table>

\[ ^{19} \text{In principle, the three parameter lognormal distribution may be applied, though this is not an option in SAS. The symmetrical nature of this distribution does not encourage a search for the three-parameter lognormal distribution.} \]
7.2 The Copenhagen-Ringsted case

Figure 7.33: Total public transport times

Figure 7.34: Total in vehicle car time
moments supports that the distribution of the access/egress time is skewed, and that the PT time does not deviate form normality.

When (variants of) the logit model is applied for estimation, in fact $\mu \beta$, is estimated rather than $\beta$, where $\mu \in \mathbb{R}$ is a scale inherited from when the variance of the Gumbel error term was fixed to 1. To reject the possibility, that distribution is caused solely by different scales for different groups, scatter-plots of pairs of coefficient estimates are produced. If the variation shown in the table above is nothing but a result of scaling, the scatter-plot should be a line through the Origin (for all pairs of coefficients), with slope equal to the ratio of the two coefficients depicted in the plot. Plots are drawn for all combinations of coefficients i.e. $\beta_i \times \beta_j$ where $i < j$, $j \in 1, .., n$. Two of these plots are presented below.

In these plots the points are widely spread in the figure, that is they do not resemble a straight line; similarly for the remaining 4 plots which are not shown. Hence, the variation in the coefficients shown in figure 1 through 4 was not caused simply by the logit model scale parameter. Accordingly, there is no reason to believe that there is no variation in the coefficients.

The matrix of correlations, which is shown below (table 7.25), addresses the possible dependency between the coefficients. A fairly high correlation (-0.50) between the (group estimates of) coefficients for cost and time spent in public transport modes is indicated (values in parenthesis are $t$-values, between group variation). The correlation between the (group estimates of) coefficients for cost and access/egress time respective car time is -0.24 respective -0.28, which is an amount that cannot be neglected either. In general there is a clear tendency that the coefficients are indeed interdependent -implying that a four-dimensional simultaneous distribution should be employed. However, at this point it should be emphasised that these estimates are based on group estimates of the coefficients, which may cover up some variation as the estimates are pulled towards the mean (one common value for all individuals in a group). Similarly, as this happens for all groups this is likely to indicate a level of correlation that is higher than what is the case.

The parameters $(\zeta, \sigma)$ in the simultaneous lognormal distribution ($LN_4(\zeta, \sigma)$)
7.2 The Copenhagen-Ringsted case

Figure 7.35: Access/egress vs. public time

Figure 7.36: Cost vs. access/egres time
### Table 7.26: $\zeta, \sigma$ parameters defining the distribution of the $\beta$-s

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$\zeta$</th>
<th>Cost</th>
<th>AE time</th>
<th>Car time</th>
<th>PT time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>-4.6</td>
<td>-10.9</td>
<td>-8.5</td>
<td>-14.9</td>
<td></td>
</tr>
<tr>
<td>AE time</td>
<td>-3.4</td>
<td>-0.1</td>
<td>7.7</td>
<td>0.5</td>
<td>-0.5</td>
</tr>
<tr>
<td>Car time</td>
<td>-1.8</td>
<td>-0.1</td>
<td>0.5</td>
<td>3.9</td>
<td>1.5</td>
</tr>
<tr>
<td>PT time</td>
<td>-2.7</td>
<td>-0.2</td>
<td>-0.5</td>
<td>1.5</td>
<td>3.3</td>
</tr>
</tbody>
</table>

The elements in the (symmetric) $\sigma$-matrix are shown below in table 7.26 (Note, that the $\sigma$-matrix is not the covariance matrix). If the elements on the diagonal are compared to the $\sigma$ values in table 7.24, a substantial reduction can be seen - due to the allowance of covariance between the coefficients. This is a clear indication that possible interrelations between coefficients should not be neglected.

When the coefficients are distributed, it is obvious, that so are the values of time (VOT). These distributions are shown below in figures 7.37 and 7.38 (in Dkr/min). The VOT is the result of two (possibly different) distributions, as the VOT is the ratio of two distributed coefficients. Hence, the distribution of the coefficients themselves cannot be assessed solely from the distribution of the VOT, unless under conditions where one coefficient is known fixed, the same distribution applies for both coefficients, etc., which is not the case here. In the figures unrealistic values (wrong sign) has been truncated. The distributions are very clearly skewed towards higher values. This is just as one could expect, as the aversion towards additional travel time can become severe, more severe, extreme severe etc., corresponding to higher and higher values of time. On the other end of the scale additional travel time can matter little, even less, etc., but never become lower than zero. This pattern results in a skewed distribution of the value of time. Re-scaling the unit to Dkr/hours, the mean and variation of the VOT is in table 7.27. Note, that rule of thumb (between std.err and mean) does not apply to the VOT, as these distributions are skewed. The skewness of the distributions is very clear from the figures – the symmetrical normal distribution is far from the distribution depicted.
7.2 The Copenhagen-Ringsted case

Figure 7.37: VOT Car time

Figure 7.38: VOT Public transport time
7.2.3 Test 2, 50 groups, 10 runs

A second test based on 50 groups was formed, 10 model runs were performed. Simple statistics are shown below in table 7.28. In general the patterns from test are reproduced though the distribution is not as smooth as for test 1. Again, a skewed distribution is suggested, with correlation between the coefficients.

7.2.4 Test 3, 25 groups, 10 runs

Test 3 is based on just 25 groups. Even in this case the distributions are clearly skewed, coefficients are clearly correlated to the same extent as for test 1.

7.2.5 All tests

As hypothesised in section 6.3.1 the number of groups do have an impact on the distribution, but fortunately, the shape and location is by and large unaltered by the number of groups. The main difference is the smoothness of the curve.
7.2 The Copenhagen-Ringsted case

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>ζ</th>
<th>σ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(10^{-1})</td>
<td>(10^{-3})</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td>0.16</td>
<td>0.11</td>
<td>0.01</td>
<td>-0.59</td>
<td>-4.3</td>
<td>0.6</td>
</tr>
<tr>
<td>Access/ Egress time</td>
<td>0.32</td>
<td>0.07</td>
<td>-1.91</td>
<td>6.22</td>
<td>-3.5</td>
<td>0.3</td>
</tr>
<tr>
<td>Car travel time</td>
<td>1.53</td>
<td>1.29</td>
<td>-2.56</td>
<td>9.69</td>
<td>-1.9</td>
<td>0.2</td>
</tr>
<tr>
<td>Public transport time</td>
<td>0.61</td>
<td>0.19</td>
<td>-3.13</td>
<td>12.28</td>
<td>-2.8</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table 7.30: Results from test 3

<table>
<thead>
<tr>
<th></th>
<th>Cost</th>
<th>AE time</th>
<th>Car time</th>
<th>PT time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>1.00</td>
<td>0.30</td>
<td>0.13</td>
<td>0.00</td>
</tr>
<tr>
<td>AE time</td>
<td>0.30</td>
<td>1.00</td>
<td>0.59</td>
<td>0.59</td>
</tr>
<tr>
<td>Car time</td>
<td>0.13</td>
<td>0.59</td>
<td>1.00</td>
<td>0.82</td>
</tr>
<tr>
<td>PT time</td>
<td>0.00</td>
<td>0.59</td>
<td>0.82</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 7.31: (Pearson) Correlation of the coefficients, test 3

For test 1, 30 runs (different groups) were carried out, partly to enable smoothing over 30 runs, and partly to enable comparisons of ‘bundles’ (10 runs) of different runs. These showed remarkably similar distributional patterns (The figures are in Appendix C).

For this case it seemed as if 10 model runs with 25 groups were sufficient for identification of the type of distribution (uni-modal, skewed) whereas determination of the distribution parameters required 10 model runs with a minimum of 50 groups (with a total run time of less than 4 hours for 18000 observations).

7.2.6 Stability of results

As mentioned earlier, the first test was based on 30 repetitions, whereas the second and third were based on 10 runs (as for the synthetic data). To assess whether this increased number of runs had any impact on the distribution, comparative plots were produced. In these plots, the distribution for bundles of 10 runs (run 1-10, 11-20 and 21-30), are presented parallel to each other. The charts are shown in figures 7.39 through 7.42.

Between the distributions of bundles of runs, there are small differences, but these are small! The shapes of the distributions are the same, the locations and the spreads are identical (axes are identical, though this may be hard to read.).

7.2.7 Conclusions of CRM data

In the original estimation of the Copenhagen-Ringsted Model the method of maximum simulated likelihood (MSL) was applied for the determination of dis-
Figure 7.39: Access and egress time

Figure 7.40: Cost
7.2 The Copenhagen-Ringsted case

Figure 7.41: Total public transport time

Figure 7.42: Total in vehicle car time
distribution of the coefficients in a mixed logit model. However, it was debated whether the 'correct' distributions were actually found, as no formal tests of the type of distribution was carried out except for comparisons of likelihood value for MNL, NL and MSL (Normal distribution). An error component was added to a 'generalised time component', implying perfect correlation between (potential) error components for each time component. Neither this was tested.

This thesis has described a new method to assess the empirical distribution of error components, as opposed to the method of maximum simulated likelihood, which assumes that the shape of the distribution is specified \textit{a priori}. The assessment of distribution involves estimating separate models for fractions of data and combining results from several model estimations. The method is without crude assumptions on which distributions to identify, on the contrary to the MSL, that 'just' optimises on parameter values given a specific shape of distribution. Furthermore, independence is not assumed, that is, the distribution of the coefficients may be constituted by a simultaneous distribution. Tests for distribution were carried out on Stated Preference data from the Copenhagen-Ringsted Model.

Neither of the coefficients in the model could be identified as being fixed (zero variance). Two of the (marginal) distributions uncovered in this test seemed to be skewed to the right. Accordingly, normal distributed coefficients will often not be the best choice (beside the well known problem of truncation, when using the normal distribution for sign restricted coefficients).

Empirical evidence was found that supports the hypothesis of correlation between different error components. Accordingly, the coefficients may be correlated in a more complex pattern than often assumed when designing error-component models.

The distributions depicted could not be explained simply as differences in the logit scale parameter \( \mu \), as scatter plots of the coefficients did not show such a linear relationship (proportionality between the coefficients \( \beta_i \) and \( \beta_j \) for \( i < j \)).

In general, data groups should be created by completely random generation – thereby preventing introduction of systematic variation in the plots of empirical distributions. Another issue, is how the number of data sets in a model affects the distributions of the coefficients, as well as how the determination of type of distribution can be automatised. A combination of graphical and (statistical) moment assessment was used.

An additional issue for future research is the functional form of the utility function. All tests were based on a simple linear form, whereas it is likely that other specifications may prove better. Including both non-linear terms, error components on coefficients for linear and non-linear elements and even correlation between these, will severely complicate the model estimation – and in particular
model forecasting.

Summarised, the empirical evidence in this test points towards a joint distribution of dimension equal to the number of attributes in the model. The type of distribution was likely to be lognormal distributed for *access/egress time* and *cartime*, though normal distributed for *cost* and *PT time* (travel time in public transport modes). However, the analysis should be repeated for other data sources than this Danish sample, as distributional patterns are expected to be subject to the context.
The offset of this Ph.D. study was to gain an in dept understanding of individual’s behaviour in relation to choice of whether to perform a trip, when to undertake it, by which mode etc. By understanding these mechanisms – or at least, be able to model them in a statistical sense, predictions on future travel patterns becomes within reach. Traffic forecasts of passenger travel is not a new area, but room for improvement is there. This was recently demonstrated in Denmark, as the Øresund Traffic Forecast Model, slightly understated, did not match the actual traffic realised by the opening of the bridge between Denmark and Sweden. Neither did it correctly predict traffic one year after the opening.

A previous major model development in Denmark, the Copenhagen Ringsted Model Complex, included cutting edge model developments but the urge for gaining further understanding of capabilities and/or limitation, was not to be neglected. Parallel to that study, the corner stones to this Ph.D. study were formulated.

The intention of this thesis was to give an overview of what has been done in the research area of passenger transport modelling, with a focus on the model type in the core of a model complex. The core of a model is occasionally nicknamed a 'black box' – hopefully, this thesis has uncovered that it may be tough to get rid of that reputation, as model are growing more and more complex.

Though, the literature reveals examples of the opposite, the starting point of any descriptive or modelling study should be description of the context. As such, this thesis commence by describing the type of the problem, that is the discrete choice between a finite set of alternatives and characteristics of these. An intuitive explanation of the maximum likelihood estimation method leads into a broader overview of the model variants proposed in the literature. This literature review of alternative proposed and implemented models, spans from model specifications with independent alternatives to more complex specifications allowing for mode characteristics to be shared by more than one alternative.
The incompleteness of models ability to exactly capture the choice formation in the human brain (the most complex neural network) is patched up by adding an error term to capture this. Unfortunately, to make models manageable simplifying restrictions has been added. In later developed models, such restrictions have been relaxed as computer power allowed for this. Models based on simple orthogonal error structures, which facilitate fairly simple estimation routines to in principle full covariance structure are covered. The review described the development within the models capabilities to include tree structures to accommodate estimation of models, where alternatives are not independent e.g. the split between different service levels for train.

An important breakthrough came in 1986 by McFadden and Train, showing that any (GEV) model capable of handling overlapping characteristics between alternatives (tree structures), could be approximated closely by use of stochastically distributed terms in the utility function. This, for the sake of theoretical model improvement, caused focus to be shifted towards such models. This result has also affected the work during this study. However, this does not mean that models based on tree structures should be abandoned! In terms of presenting a model to a layman, a colleague, or getting a quick grasp of how a model is structured tree models are magnificent.

Relations between the models covered in this section, that is which are a generalisation of which, are shown graphically. Further, the models are divided into three classes, namely Hybrid choice models, Tree models and Latent class models. Hybrid choice models covers all models where at least on additional distributed term is included (compared to the residual). Tree models covers all models where the model structure is hierarchical, that is, alternatives may be grouped, sub-grouped, further sub-grouped, and so on. Models where cross-nesting between alternatives is allowed, are also included here. Both these classes of models allows for dependencies between alternatives. As the name indicates, the final class, Latent class models, covers models where individual choice (or part of their choice) can be described by latent (unobservable) classes, which each contribute to describe the choice. Membership of the latent classes may be described/ modelled by probability density functions.

The missing link – though, fortunately not any longer, between computationally complex stochastic models and estimation results has been facilitated by developments in computer power. The mean is simulation and the goal in infinitesimal close approximation of complex model formulations. The closing of the literature review briefly describes traditional iterative methods for model estimation followed by simulation methods used for the stochastic models. This chapter was included as improvements in discrete choice models cannot stand alone. No matter how ‘perfect’ a model is, it cannot patch up for poor data quality, nor for an inaccurate model estimation (the optimisation procedure). With the current speed of development within optimisation routines, recent findings
should be consulted if one is interested in the most appropriate optimisation method for a given problem.

In the thesis, special emphasis has been put on two issues, namely, data segmentation and mixed logit models. Recognition of differences between individuals, that is different (relative) preferences, may be handled by segmentation of the data the model is based upon. In reality, different models are estimated from subsets of the data taking advantage of, after all, having some (assumed) consistent data to work from. Data segmentation has previously been based on 'sound knowledge', e.g. segmentation by trip purpose. In the thesis a new method was proposed, which segments data accordingly to individual's (estimated) preferences. Segments are formed such that individuals homogeneous in preferences belong to the same segment whereas individuals with different preferences are in different segments. The method is a heuristic, in part because it is based on heuristic methods for identification of the segments, and partly because data limitations (generally too few observations per individual) has been overcome by pooling individuals, without the option of testing for homogeneity of the pooled individuals.

A small case was discussed, though the case was 'too small' to determine the worth of the method. The method revealed a potential for improvement of the fit of model to data, compared to both an un-segmented model and a purpose-segmented model. Three models with alternative segments were constructed - two of these were significantly better than the purpose-segmented model. For the present data set, a short fall of the method was illuminated. The segments are constructed be means of cluster analysis; the last step of the method is a translation of the clusters into segments. Whether this translation is possible depends on the data as segments are to be described by variables (socioeconomic, individual specific, etc.) in the data. In this case, it was not possible to recover a simple description of the segments based on the variables in the data set (only very few variables were available).

The second proposed method related to the estimation of Mixed Logit models. The most commonly used approach to estimate Mixed Logit models is to employ the Maximum Simulated Likelihood estimation (MSL). The MSL finds optimal coefficients values of ordinary utility elements simultaneously with optimal parameter values for distributed terms in the utility function. Like ordinary maximum likelihood estimation, the utility function must be specified prior to the estimation as a function of some attributes and distribution parameters to be estimated. In particular, for all included distributed terms the shape of the distribution must be specified.

The proposed method, assesses the shape of the distribution from data, by means of repetitive model estimation. In this case, a model was estimated based on partitioned data, where one model was estimated for each sub-sample and repetitions reflects different partitions of the data set. The shape of coefficient
distribution is assessed from between model comparisons. Again the method is a heuristic, since data limitations (generally too few observations per individual) has been overcome by pooling individuals.

The proposed method for assessment of shape of distribution is not to be regarded as an alternative to MSL estimation, rather as a complimentary test to assess the shape of distribution, which is to be assumed a priori in the MSL modelling.

The two proposed methods, are related in the sense that they both rely on successive model estimation from grouped (partitioned) data. The difference is the formation of the groups and the subsequent treatment of the coefficient estimates. For the segmentation test, groups are generated such that records (individuals) are similar in terms of one (some) variable(s) to aid in the later identification of segments. For the distribution assessment, the groups are formed by random sampling to avoid conditioning the estimation results (and thereby shape of distributions) on some additional information. The subsequent treatment (or inter-model comparisons of results), for the segmentation test consists of finding patterns in the coefficient estimates like combinations of above/below some threshold value for some/each of the coefficients. For the distribution assessment, a search for shape follows, that is, differences between groups coefficient estimates are used to create the density chart for the coefficient. At this point, it is clear that Mixed Logit models may be seen as a continuous variant of data segmentation.

The application section for the distribution assessment was founded on two data sources. Synthetic data was developed to test the method, after which it was employed on empirical data; this data had previously been validated in conjunction with a model development project in Denmark. The synthetic data set demonstrated a problem with the way MSL models are constructed/validated, as the $t$-tests showed significance of variance term even for mis-specified shape of distributed terms! This indicates that even mal-specified variance term contribute in explaining some of the total variance in the data, but more importantly, that it cannot be used as an indicator for whether the correct distribution has been specified in the model! The $t$-test has traditionally been consulted when a model was tested, wherefore a number of models reported in the literature, may or may not, be erroneously specified. Generally, normal and lognormal specified variance terms were significant irrespective of the 'true' shape of distribution. Using a combination of direct comparisons of the likelihood value and the $t$-statistic, did aid in limiting the plausible specifications for the variance terms.

The proposed method for assessing shape of distribution had some problems in recovering the correct shape, in the cases where only one attribute described the choice of alternative. For the choice data based on two attributes, the results were very promising – the two different shapes (one normal, one lognormal)
were recovered correctly and with a correct indication of which coefficient had the highest variance. Further tests on synthetic data, with choice based on at least two attributes were recommended.

On the empirical data, distributions were assessed by the method just validated on synthetic data. Compared to the original model, the model was a simplified version based only on four attributes. In the test, distribution was allowed for all coefficients – and it turned out that variation could be found for all coefficients. The shape of distributions was different, though all were smooth, and uni-modal. Two of the coefficients’ distributions were symmetrical and two were skewed – or more specifically, two (Cost and PT time) were similar in shape to the normal distribution, whereas the remaining two (Car time and access/egress time) were similar in shape to the lognormal distribution. Since the method revealed both skewed and non-skewed distributions just as the final synthetic data set, it is plausible (though not justified in a statistical sense) the ‘correct’ shape has been recovered. Furthermore, the distribution was not caused by the logit scale (this was tested separately).

This Ph.D. thesis has revealed, that though the theory of traffic modelling has improved in a remarkable pace over the past 40 years, there is still room for further research. The matters of segmentation and mixed models, may be regarded as two sides of the same question, as both recognise that individuals are different – but to a varying extent. The models based on segmentation are far simpler than the models involving distributed terms, which justify their application within simpler modelling studies. However, proof of existence of segments that do improve models, or at least, a method to prove such, and proof of segments that are supported by empirical evidence, is lacking in the literature.

The area of shape of distribution is, by my opinion, just beginning to unfold, more research into shape, and probably multi-dimensional (non-independent) distributions are needed. Likewise a tool for determining the significance of the shape of a distribution, should be developed, as the t-test only indicate whether a variance term contribute to explain some of the variance in the data, and not indicate significance of the shape of the distribution! Provided, that we can describe behaviour by complex utility functions, a ‘translation’ into a tree structure, to aid in explaining the models to a layman, is also on my wish list.
Appendix A

Statistics

A.1 Moments

Moments (around the mean) of a variable $X = X_1, \ldots, X_N$ in the sample $S$, are specified as

$$\hat{m}_k = \frac{\sum_{j=1}^{N} (X - \bar{X})^k}{N} \quad (A.1)$$

The relationship between estimates for disjoint exclusive groups $g_1, \ldots, g_G$ where $\bigcup g_j = S$ and $g_l \cap g_j = \emptyset, \forall \ l, j$. Groups are formed by random sampling and are equally sized (size $N_g$). Estimators for groups are

$$\hat{E}(X_j) = \frac{\sum_{j=1}^{N_g} X_j}{N_g} = \bar{X}_g \quad (A.2)$$

$$\hat{Var}(X_j) = \frac{\sum_{j=1}^{N_g} (X - \bar{X}_g)^2}{N_g}$$

$$\hat{m}_k = \frac{\sum_{j=1}^{N_g} (X - \bar{X}_g)^k}{N_g}$$

Estimators for the sample can be calculated from the estimators of the groups as

$$\hat{E}(X_j) = \frac{\sum_{j=1}^{N} X_j}{N} = \sum_{g=1}^{G} \bar{X}_g \frac{N}{N_g} \quad (A.3)$$

as the groups are formed by random sampling. The mean value is the average of the group mean values corrected for differences in group size. The second
that is, the group variances are multiplied by the size of the groups. The third moment (skewness) is determined as

$$\hat{m}_3 = \frac{\sum_{j=1}^{N} (X_j - \bar{X})^3}{N}$$  \hspace{1cm} (A.5)

A.1.1 Moments calculated from sub-samples

Moments to account for the variation within groups as well as between groups can be estimated from the following formulae:

$$s^2 = \frac{\sum_g N_g \sigma_g^2}{\sum_g N_g} \hspace{1cm} (A.6)$$

$$\hat{m}_3 = \frac{\sum_g N_g m_{3(g)}}{\sum_g N_g} = \frac{\sum_g N_g m_{3(g)}}{N} \hspace{1cm} (A.7)$$

where $N_g$, $m_{3(g)}$ are number of observations resp. estimate of third moment for group $g$. Estimate of moments are not readily available for groups; these can be approximated by use of Jackknife estimation. That is, estimates $\tilde{\theta}$ of an estimator $\theta$ are obtained by

$$\tilde{\theta} = n\hat{\theta} - (n - 1)\frac{\sum_i \hat{\theta}_i}{n}$$  \hspace{1cm} (A.8)

where $n$ is number of elements, $\hat{\theta}$ is estimate based on $n - 1$ observations. Specifically, variance is found as

$$\hat{\sigma}_J^2 = \frac{n-1}{n} \sum_i (\hat{\theta}_i - \bar{\theta})^2, \bar{\theta} = \frac{1}{n} \sum_i \hat{\theta}_i$$  \hspace{1cm} (A.9)
A.1.2 Correlation

Typically by correlation we mean linear correlation measured from a sample \((r)\) or a population \((\rho)\) as

\[
\hat{r}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}.
\] (A.10)

Significance of correlation can be tested by the simple test for sample correlation given by

\[
t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}},
\] (A.11)

where \(r\) is the sample correlation based on \(N\) observations in the sample. The test has a Student’s distribution with \(\nu = N-2\) degrees of freedom (given that \(\rho = 0\)).

A confidence interval can be determined by use of Fisher’s Z-transformation, which is generally used to test for hypotheses like \(\rho = \rho_0 \neq 0\). For a description of correlation measures see e.g. (Spiegel, 1998).

A.2 Gumbel (or Extreme Value, Type I) distribution

The Extreme Value, Type I (EV1) distribution is also referred to as the Gumbel distribution (and by some, even erroneous by the Weibull distribution). The distribution function is skewed compared to the Normal distribution, see figure A.1.

The explanation for the emphasis on 'Type I' for the distribution is, that there are three different extreme value distributions. The distribution functions are different, as the range they are defined on are different, see table A.1. The two-parameter distribution function is given by

\[
F(x, \alpha, \beta) = \exp\left(-e^{-\left(\frac{x-\alpha}{\beta}\right)}\right), \text{ with }
\] (A.12)

\[-\infty < x < \infty, -\infty < \alpha < \infty, \beta > 0,
\]

with the density function described as

\[
f(x, \alpha, \beta) = \frac{1}{\beta} \exp\left(-e^{-\left(\frac{x-\alpha}{\beta}\right)} - \left(\frac{x-\alpha}{\beta}\right)\right)
\] (A.13)

where \(-\infty < x < \infty, -\infty < \alpha < \infty, \beta > 0\).

Mean and variance may be calculated by

\[
E(x) = \alpha + \gamma/\beta,
\] (A.14)
Figure A.1: Density of the Extreme Value, Type I distribution compared to the Normal distribution

<table>
<thead>
<tr>
<th>Type</th>
<th>Distribution function</th>
<th>Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>( \exp(-e^{-x}) )</td>
<td>(-\infty \leq x \leq \infty)</td>
</tr>
<tr>
<td>Type II</td>
<td>0</td>
<td>( x \leq 0 )</td>
</tr>
<tr>
<td></td>
<td>( \exp(-e^{-\alpha}) )</td>
<td>( x \geq 0 )</td>
</tr>
<tr>
<td>Type III</td>
<td>( \exp(-(e^{-\alpha}) )</td>
<td>( x \leq 0 )</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>( x \geq 0 )</td>
</tr>
</tbody>
</table>

Table A.1: Density functions for Extreme Value distributions.
A.3 Log-normal distribution

where \( \gamma (= 0.577) \) is Euler’s constant, and

\[
\text{var}(x) = \frac{\pi^2 \beta^2}{6}
\]  \hspace{1cm} (A.15)

Other characteristics of the Extreme value distribution are

- If \( X \sim \text{EV}_1 \), then \( Y = \exp X \) is Weibull distributed.
- For fixed \( \alpha \) and \( \beta \), the density function \( f(x, \alpha, \beta) \) has 'mode' in \( x = \alpha \).
- If \( X \sim \text{EV}_1 \) with density function \( f(x, \alpha, \beta) \), then the largest order variable \( Y = X_{n,n} \) is distributed \( X_{n,n} \sim \text{EV} \), with density \( f(y, \alpha, (\beta/n)) \).

The last property ensures, that is we remove the best alternative, the residual is still distributed EV1.

A.3 Log-normal distribution

If a stochastic variable \( Y \) is normal distributed, then \( X = \exp(Y) \) is log-normal distributed. The probability density function for a log-normal distributed variable \( X \sim \text{LN}(\zeta, \sigma) \), is given by:

\[
P_X(x) = [x\sqrt{2\pi\sigma}]^{-1}\exp[-1/2(\log(x) - \zeta)^2/\sigma^2] \text{ for } x > 0.
\]  \hspace{1cm} (A.16)

Estimates of the distribution parameters can be calculated from the estimates of the mean and the variance of the coefficients. These are given by

\[
\text{Mean} = e^\zeta \omega^{1/2} \text{ where } \omega = \exp(\sigma^2) \text{ and } \quad (A.17)
\]
\[
\text{Variance} = e^{2\zeta} \omega (\omega - 1) \text{, hence } \quad (A.18)
\]
\[
\text{Coefficient of variation} = \sqrt{(\omega - 1)}.
\]

For simultaneous distributions the parameters \( (\zeta, \sigma) \) can be calculated from the mean and the covariance, given by

\[
cov(X_i, X_j) = \exp(\zeta_{X_i} + \zeta_{X_j} + 1/2\sigma_{X_i}^2)(\exp(2\sigma_{X_{i,j}}) - 1) \quad (A.19)
\]
\[
= E[X_i]E[X_j](\exp(2\sigma_{X_{i,j}}) - 1),
\]

where \( \zeta = \begin{pmatrix} \zeta_{X_1} \\ \zeta_{X_2} \\ \zeta_{X_3} \\ \zeta_{X_4} \end{pmatrix} \) and \( \sigma = \begin{pmatrix} \sigma_{X_1}^2 \\ \sigma_{X_2}^2 \\ \sigma_{X_3}^2 \\ \sigma_{X_4}^2 \end{pmatrix} \).
A.4 Test for distribution

A well-known distribution-free test for tests for shape of distribution is the Kolmogorov-Smirnov test K-S. The K-S test can be applied to test whether data follow any specified distribution. As a general test, it is not as powerful as a test designed specifically to test for normality. Moreover, the K-S test loses power if the mean and/or variance is not pre-specified. The K-S test will not indicate the type of non-normality e.g. whether the distribution appears to be skewed or heavy-tailed (kurtosis). The K-S test is defined as

\[ D = \max_{1 \leq i \leq N} |F(Y_i) - \frac{i}{N}| \]  

(A.20)

The Shapiro-Wilk (S-W) test for normality determines whether or not a random sample of values follows a normal distribution. The test is designed to detect departures from normality, without requiring that the mean or variance of the hypothesized normal distribution be specified in advance. The test tends to be more powerful than the K-S test, though it will not indicate the type of non-normality (skewed, heavy-tailed or both). The S-W test is calculated as

\[ W = \left( \sum a_i x(i) \right)^2 / \left( \sum (x_i - \bar{x})^2 \right) \]  

(A.21)

where \( x(i) \) is the \( i \)th largest order statistic, \( \bar{x} \) the sample mean, \( a_i \) coefficients. Small values of \( W \) are evidence of departure from normality. The standard algorithms for the S-W test only apply to sample sizes up to 2000. For larger samples Stephens’ test for normality can be applied.

The Anderson-Darling test \( (A^2) \) is used to test if a sample of data follows a specific distribution. It is a modification of the K-S test and gives more weight to the tails of the distribution than the K-S test. The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution. The Anderson-Darling is defined,

\[ A^2 = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \left( F(x)(1 - F(x)) \right)^{-1} dF(x) \]  

(A.22)
Appendix B

Alternative segmentation of CRM data

When the Copenhagen Ringsted model was constructed, the model was formed based on the available data, re-using existing code and to some extent re-using existing segmentation of data. Individuals were assigned to segments depending on the purpose of their trip, as it has been recognised that individuals travelling with the purpose ‘to/from work’ generally can be described by a set of relative preferences different from an individual doing a shopping trip. Models formed on un-segmented data as well as for each of the purpose-defined segments, were compared in the light of model fit (likelihood) and coefficient values (based on sound knowledge).

Subsequently, a new method for segmentation of data was proposed, the method was applied to the data from the Copenhagen-Ringsted model. The method was applied to the RP part of the data. The data consist of 3 RP data sets consisting of 6700 records. In this case each observation corresponds to one individual performing one trip and only one observation for each individual, and therefore no potential repeated measurements error to account for. For a description of the data and/or the model see Nielsen (2000). Note, that the data applied in this section (RP data) is not identical to the data applied in the previous section (SP data).

B.1 Results

In the following, a number of models are set up; initially the traditional MNL model with no segmentation (for reference), two different purpose-segmented model and finally three different models resulting from the application of the suggested methods. The model setup is kept constant for all models, that is the same variables are included in the utility function.
Model 1: No segmentation
This model was constructed to have a basis for evaluations of the following segmented models. The Likelihood value was $L^1 = -49,053.71$; the model had five coefficients.

Model 2: Segmented by purpose, version 1
Traditionally traffic modelling has been segmented by purpose of the trip. In this data set, purpose was specified in three categories; home-work/education, Business and Other. The obvious advantage of this segmentation was that it was easily interpretable and therefore more easily transferable to other applications, eg. strategic planning. Moreover, the segmentation is simple (requires only one variable to define segments). The Likelihood value for the model was $L^2 = -43,308.87$. Compared to the unsegmented model this is superior, as the likelihood ratio test ($LR_{(df)} = -2 \cdot \log(L^R - L^U)$ is 0.000097 with 10 degrees of freedom (df).

Model 3: Segmented by purpose, version 2
A further subdivision of the home-work/education and the 'other' segments resulted in five purposes. This lead to a likelihood of $L^3 = -41,600.42$. Compared to the model with three purpose segments this seems superior, and the $LR_{(10)} = 0.00000$ (comparison with the unsegmented model), though, the number of observations in the 'new' segments are quite small.

Model 4: Application of algorithm, version 1
This analysis follows the above described algorithm. Initially 50 groups were constructed from data; specified by the variables origin, destination and purpose. The smallest respective the largest group had 95 respective 195 observations. Coefficients estimated for all 50 groups (250 coefficients in total) were used as input for the cluster analysis. The objective of the cluster analysis was to identify which records were 'close' to which other records of data. The 'closeness' was measured by distance between records or as clusters are formed, between clusters. A total of five different distance measures was used, as neither one of those had been proven superior to all the others. These measures are based on distance between cluster (centroids or outermost observation), $R^2$ or the likelihood value of the clustering. A view of how the actual clustering was performed is shown in figure B.1 below. The X-axis indicate the records (or as here, groups of records), while the Y-axis measures 'distance' between clusters, ie. the higher the value the more these clusters differ in coefficient values $\beta$-s. Or stated differently, the higher up in the tree the ramification takes place, the stronger the segmentation is.

The model was re-estimated based on the new coefficients. The total likelihood was $L^4 = -42,760.51$, which is quite an improvement over the unsegmented
Figure B.1: Clustering for model 4
APPENDIX B. ALTERNATIVE SEGMENTATION OF CRM DATA

model (1), \( LR_{(30)} \) is 0.00000), furthermore, it is seems better than the purpose-segmented model, although, these models are not directly comparable (they are not hierarchical).

Model 5: Application of algorithm, version 2

The number of starting groups is 100 based on the same criteria as above in order to describe the effect of the number of starting groups. The smallest respective the largest group had 47 respective 95 observations. These are clustered by their coefficient estimates into 10 clusters. The total likelihood is \( \mathcal{L}^5 = -40,559.39 \), corresponding to \( LR_{(45)} \) is 0.00000.

Model 6: Application of algorithm, version 3

The previous model ended up with 10 segments, which is quite high given the amount of data available (even though the model was better than the unsegmented model in terms of likelihood ratio). To overcome this problem the algorithm was re-applied to the data. As before the number of starting groups is 100 based on the same criteria. The final model with three clusters has a total likelihood is \( \mathcal{L}^6 = -48,194.40 \). This is less than the likelihood of the unsegmented model, with a larger number of coefficients! Hence, this model is clearly rejected.

The final step of the algorithm, depicting the segments turned out problematic with the applied data set. The data contained only few socioeconomic variables which could not be used for description of segments. Hence, evidence on how the likelihood is affected by the final stage cannot be presented.

B.2 Summary

This section show that it was possible to construct alternative segments (clusters) and apply these to traffic demand models. The results show rather large differences in performance of these clustering’s as well as a need for optimisation of which characteristics the initial groups should be based upon as the resulting segments are dependent on these. In table B.2, the results are summarised.

The unsegmented model is rejected compared to model 2 through 5 by the likelihood ratio test; model 6 is rejected compared to model 1 (lower likelihood value).

Obviously the estimation time has increased due to the additional test of segments - with respect to the improvement of the model fit this seems worthwhile. The test of segments needs to be coordinated with the demand modelling as some variables, which in the demand model were insignificant (significant), may turn out being significant (insignificant) in the segmented model, simply due to the division of the data.
B.2 Summary

<table>
<thead>
<tr>
<th>Test</th>
<th>Log likelihood</th>
<th>Description</th>
<th># Coeff</th>
<th>Identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-49,063.71</td>
<td>Unsegmented</td>
<td>5</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>-41,600.42</td>
<td>5 purposes</td>
<td>25</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>-43,308.87</td>
<td>3 purposes</td>
<td>15</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>-42,760.51</td>
<td>7 segments based on 50 groups</td>
<td>35</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>-40,559.39</td>
<td>10 segments based on 100 groups</td>
<td>50</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>-48,194.40</td>
<td>3 segments based on 100 groups</td>
<td>50</td>
<td>No</td>
</tr>
</tbody>
</table>

Table B.1: Results from segmentation

Alternative segmentations can be applied throughout the sector of transport supply, to target new products or to aim a campaign directly at those who may benefit from it - thereby reducing the cost of advertising. In general, the method is sensitive towards the data to which it is applied. To the knowledge of the author, no similar study has been undertaken; therefore there is no other (alternative) segmentations to compare with.

There is need for further work within this area, in order to secure new guidelines for how to segment data for use in traffic analyses. The first step is to apply the method to another data set to verify whether the here experienced problems with description of segments persist. Of not, issues of improving the model are relevant. These include re-running the segmentation to find segmentations that are stable over different partitions of data and investigate hierarchical segmentation structures in more detail. Stability of segments over geography/time is another issue, which may be asserted based on comparison of a number of different data sources (different areas) preferably from different time periods.

The ongoing development of error component models may be seen as another indication of the need for describing individuals by different coefficients. In fact, the models with random coefficients may patch up for segmentation not included in the models. An advantage of segmentation over models with random coefficients is the lower estimation time for the segmented model. The difference between the two approaches is rather small as the estimation of coefficients from grouped data can be applied for assessing the distribution of the coefficients, see chapter 6. A second explanation for the urge of use of segmented model rather than mixed models is, that it is immediately intelligible for a planner (and much easier to present to a layman), whereas mixed models may be more difficult to comprehend, especially when is come to matters of shape of distribution.
APPENDIX B. ALTERNATIVE SEGMENTATION OF CRM DATA
Appendix C

Figures for distributed coefficients
APPENDIX C. FIGURES FOR DISTRIBUTED COEFFICIENTS

Figure C.1: Distribution of $\beta$ for N, 50 groups

Figure C.2: Distribution of $\beta$ for LN, 50 groups
Figure C.3: Distribution of $\beta$ for F, 50 groups

Figure C.4: Distribution of $\beta$ for ND, 50 groups
Figure C.5: Distribution of $\beta$ for NLN1, 50 groups

Figure C.6: Distribution of $\beta_2$ for NLN2, 50 groups
Curves: Normal (Mu = 0.8416, Sigma = 0.2936) Lognormal (Theta = 0, Shape = 0.3, Scale = -0.2)

Figure C.7: Distribution of \( \beta \) for N, 250 groups

Curves: Normal (Mu = 1.9779, Sigma = 0.4618) Lognormal (Theta = 0, Shape = 0.24, Scale = 0.65)

Figure C.8: Distribution of \( \beta \) for LN, 250 groups

187
APPENDIX C. FIGURES FOR DISTRIBUTED COEFFICIENTS

Figure C.9: Distribution of $\beta$ for F, 250 groups

Figure C.10: Distribution of $\beta$ for ND, 250 groups
Figure C.11: Distribution of $\beta$ for TP, 250 groups
Appendix C. Figures for Distributed Coefficients

Figure C.12: Distribution of $\beta$ for NLN1, 250 groups

Figure C.13: Distribution of $\beta_2$ for NLN2, 250 groups
Figure C.14: Distribution of $\beta$ for N, 500 groups

Figure C.15: Distribution of $\beta$ for LN, 500 groups
Figure C.16: Distribution of $\beta$ for F, 500 groups

Figure C.17: Distribution of $\beta$ for ND, 500 groups
Figure C.18: Distribution of $\beta$ for NLN1, 500 groups

Figure C.19: Distribution of $\beta_2$ for NLN2, 500 groups
APPENDIX C. FIGURES FOR DISTRIBUTED COEFFICIENTS

Figure C.20: Distribution of $\beta$ for SP, 500 groups
Bibliography


M. Ben-Akiva, D. Mcfadden, K. Train, J. Walker, C. Bhat, M. Bierlaire, D. Bolduc, A. Boersch-Supan, D. Brownstone, D.S. Bunch, A. Daly,


A. Daly, 2002. Personal communication.


Øresundskonsortiet. Traffic forecast model. the fixed link across øresund. Can be obtained by contacting Øresundsbron. Vester Søgade 10. DK-1601 Copenhagen V. E-mail: info@oeresundsbron.com, Internet: www.oeresundsbron.com, 1999.


BIBLIOGRAPHY


P. Vovsha. The cross-nested logit model: Application to mode choice in the tel-aviv metropolitan area. Presented at the 76th annual meeting of the Transportation Research Board, 1996.


