CHOICE MODELS AND BAYESIAN METHODOLOGY

AT

THE UNIVERSITY OF SYDNEY

PIERRE-FRANÇOIS ULDRY

ADT letter
sent 26/3/01

CHOICE MODELS AND BAYESIAN METHODOLOGY

At

The University of Sydney

Pierre-François Uldry

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Marketing University of Sydney

August, 2000

Abstract

The objective of this research is to develop new ways to estimate discrete choice models using Bayesian methods. Markov Chain Monte Carlo methods (more specifically, Gibbs samplers) are powerful ways to estimate many model parameters. This research focuses on discrete choice models and ways to estimate the elements of the error covariances.

Previous work using Gibbs samplers include variable selection procedures that optimally select and estimate subsets of variables in linear models. The same problem is considered here, but instead we focus on selecting elements of various covariance matrices to identify non-zero elements. This restricts the estimation to a more parsimonious structure, which should lead to a more accurate model form.

Discrete choice models can be estimated from real market and/or stated intentions data. Real market data typically are not well structured or behaved, which poses analytical challenges that require more sophisticated methods. In contrast, experimental data are better structured and behaved, which should benefit from application of more complex choice modeling techniques. Monte Carlo simulations are used to investigate situations in which true error structures violate the IID assumption of simple Multinomial Logit (MNL) models, but MNL performs as well as more complex Multinomial Probit models.

Finally, a Bayesian approach is developed and applied to combine consideration and choice information in choice experiments. The method accounts for preference heterogeneity and uses element selection procedures to simplify the error covariance matrix structure. The proposed procedure takes advantage of the additional preference information and can account for non-systematic missing data; it also can be applied to ranking data.

Acknowledgements

I wish to thank my dissertation supervisors Professor Jordan J. Louviere and Professor Robert Kohn for the time and effort they spent on my behalf.

I am thankful to Jordan Louviere for introducing me to the magnificent world of marketing research and to Robert Kohn for spending so much time teaching me the niceties of Bayesian methods. The numerous meetings and interactions with both of them guided me during the course of this study.

During my time as a PhD student, the department of Marketing has seen a great variety of professors and students many of whom I am happy to count as very good friends. Benedict Dellaert, David Bunch and recently Thomas Reutterer have all advised and helped me so much that I cannot do them justice in a few lines of acknowledgment. I have also been lucky to meet Professor David Hensher who has been very supportive and whose interesting projects provided me with a financial support. I also wish to thank other PhD candidates and friends namely Jeff Brazell, Chris Diener, Thomas Crook, John Rose and Paul Burke with whom I enjoyed interacting as well as all honour students who always brought a good ambiance to the department, particularly Eugene Chan, April Reyes and David Klein.

I am also grateful to Robert Bartels and Denzil Fiebig of the Econometrics Department who taught me during my time as a Master student at this University and encouraged me to pursue this degree. I thank my parents, Serge and Germaine, my brother Marc and my friend Fabrice Rapp for their constant support. Finally I would have never come this far without the endless help and support of Valerie Severin.

Table of Contents

Chapter 1

Introduction

Marketing strategists and/or researchers focus on consumers at micro and macro levels, depending on their strategic and/or research objectives. For example, consumer preference trade-offs might be investigated at a population level in the case of new mass-market product introductions; but other applications (e.g., e-tailers of books and CD's) may require individual level estimates to customize offerings. Both preceding cases potentially can be addressed by designing and implementing a conjoint analysis or choice experiment that permits one to quantify the trade-offs. Additionally, these techniques are used to evaluate and predict changes in market shares, competitive analyses and/or assessing pricing strategy, to name just a few other applications. Aggregate (macro) or individual (micro) levels of analysis primarily impact the estimation procedure. Indeed, some modeling approaches are better suited to uncover individual level parameters and obtain more precise insights into market structure, while others offer speed and simplicity of estimation when less detail suffices to address research objectives.

Thanks to its versatility, conjoint analysis (CA) has become a popular technique in marketing. CA relies on statistical design theory to systematically manipulate attribute levels to generate product profiles that are presented to respondents. CA's objective is to allow one to estimate the impact of each attribute independently of all others; hence, in typical applications, the attribute columns of the statistical designs used are orthogonal. Traditionally, CA respondents have been asked to rate or rank scenarios that are created by the designs; however, choice experiments have attracted growing interest and applications since they were introduced by Louviere and Woodworth (1983). Choice experiments are attractive because they can simulate real life choice situations more closely and provide more realistic and flexible prediction options than traditional CA. For these reasons, the present research focuses on designed choice experiments. Although the level of aggregation of the analysis can be changed, other differences may be introduced when constructing experiments. For example, researchers can choose whether to design attribute columns to be orthogonal a) only within-alternatives or b) within- and between-alternatives. Option (a) should allow estimation of only generic utility specification parameters, but option (b) should allow one to estimate alternative-specific utility specifications as well as choice model forms that assume non-IID errors.

The focus of this research is on choice models and is restricted to the consideration of Random Utility Models (RUM), such as Multinomial Logit (MNL) or Multinomial Probit (MNP). MNL models are widely used in marketing to analyse choice data mainly because a) estimation software is widely available, b) estimation of MNL models (typically using Maximum Likelihood procedures) is quick and easy and c) the estimated models permit fairly straightforward predictions of choice probabilities because a closed form expression exists. However, MNL is restrictive in that errors are independently and identically distributed (IID), which in the case of the Gumbel distribution implies the IIA (independence of irrelevant alternatives) property. The latter means that the odds (probability ratio) of choosing two alternatives is independent of choice set composition. Research has shown that this may be too restrictive an assumption in some marketing applications (e.g., Bucklin and Lattin 1991; Chintagunta 1993; Dellaert, Borgers and Timmermans 1996).

MNP models do not satisfy the IIA axiom, hence are more general and behaviourally appealing; unfortunately, however, MNP does not have a closed form expression for the choice probabilities. Thus, estimation of MNP models (typically via Maximum Likelihood procedures) is more complex than for simpler models like MNL because they require evaluation of multi-dimensional integrals, and for more than four choice alternatives, numerical evaluation of this integral often is very demanding. Probability simulators such as the GHK (after Geweke, Hajivassiliou, Keane) have been developed to overcome this problem.

Another approach lies in Bayesian methods, with the use of Markov Chain Monte Carlo sampling schemes. The latter have the advantage of not requiring the simulation of choice probabilities, and hence are used in this thesis to estimate MNP models. A problem with MNP models is that the number of elements of the error covariance matrix increases quadratically with the number of choice alternatives, which can lead to loss of statistical power in estimating large covariance matrices, especially given conventional sample sizes used in choice experiments (e.g., 400 subjects x 16 choice sets).

A simple solution to this problem would be to collect additional preference information. That is, most choice tasks only ask respondents to select their most preferred option, but research demonstrates that additional stated preference questions can add statistical information, even if marginally (e.g. Chapman and Staelin 1982; Hajivassiliou and Ruud 1994). This can be especially beneficial if respondents choose a constant alternative such as "none" or "current brand" as their preferred option with high frequency, because these choices provide no information about preference trade-offs per se. Thus, if one adds questions like "which of these options would you seriously consider?," one can obtain extra information about trade-offs that results in more accurate estimates.

Thus, following the discussion, this research first exploits the benefits of Bayesian methodology, and more specifically, selection procedures to estimate MNP models that select an optimal, and ideally more parsimonious, covariance structure. Bayesian selection procedures can be viewed as a Bayesian version of frequentists' stepwise selection methods. Bayesian selection procedures are data driven, so choice of a parsimonious structure does not require researchers to make unnecessary assumptions that might affect model results. This line of research is interesting because it leads to models that allow very general error covariance structures that may nevertheless converge to very simple ones comparable to the like of MNL models.

This is not to say that more general models are necessarily better than the simple ones like MNL, especially in the case of experimental data that are designed to yield the best estimates possible. Indeed, it may well be the case that, depending on one's research objectives, MNP models may not outperform MNL for some types of choice experiments. Thus, a second objective is to investigate whether MNL models may not be sufficient with high-quality data and whether certain types of designed experiments can yield sufficiently rich information to estimate the full error structure assumed by MNP. Specifically, the conditions under which MNP models perform well are not yet well-understood; hence it is important to investigate which design strategies work best in conjunction with certain model types.

Finally, individuals' preferences and tradeoffs may differ; hence a distribution of heterogeneity can be assumed/imposed to reflect and account for this uncertainty and obtain more precise representations of market structure. In fact, one can estimate individual-level coefficients, but as numbers of attributes increase, models that impose this type of structure require increasingly larger covariance matrices. Hence, the present research uses Bayesian selection procedures to find more parsimonious representations of covariance matrices. This technique is appealing because Bayesian methods permit one to estimate models that can decide on a general or a basic covariance structure depending on the data. Moreover if two (or more) structures are likely, estimates are obtained using a weighted average of the two (or more) models. Model averaging concepts can be used in many applications; this study focuses on MNP models because of their growing popularity in marketing research. Moreover using additional preference information (e.g., "consideration" responses) is proposed to estimate MNP models that take preference heterogeneity into account in order to obtain better estimates of model parameters.

The next chapter reviews conjoint analysis and discrete choice models. Chapter 3 develops a new Markov Chain Monte Carlo (MCMC) sampling scheme for MNP models that can be used to select relevant elements of the covariance matrix; that is, the procedure allows the identification of non-zero elements to obtain a more parsimonious structure. Chapter 4 considers the choice of using a MNL or MNP model to analyse experimental data. Experimental data are designed to have better statistical properties than real market data; hence, one may ask whether one needs to use more complex models (such as the one presented in Chapter 3) with such data. The research presented in Chapter 4 attempts to answer this question, and investigates whether more complex experimental designs are useful in conjunction with MNP models. Chapter 5 proposes an individual parameter model, which applies the selection procedure of Chapter 3 to the elements of the covariance matrix of the errors as well as to the covariance matrix of the individual parameters. Thus, it accommodates a parsimonious covariance structure on the individual preferences as well as a parsimonious representation of the covariance structure of the utilities. In addition, the model allows one to combine choice and more in-depth preference response data such as considerations or ranks, and efficiently accounts for missing information at random (Little and Rubin 1987). This manuscript ends with a discussion and conclusions in Chapter 6.

Chapter 2

Methodological Background

2.1 Conjoint choice experiments and Choice models

Conjoint measurement was first proposed by Luce and Tukey (1964) and introduced to marketing by Green and Rao (1971). Extensions of this approach have grown rapidly in popularity in past decades and are now widely used in academic and commercial applications. Conjoint analysis is useful to identify new product concepts, pricing strategies, market segments, analyse competitors and reposition products (see Cattin & Wittink 1989; Green and Srinivasan 1990). Recently, commercial software has become available to design and implement conjoint studies (e.g., ACA developed by Richard Johnson of Sawtooth Software), making the technique accessible to a wider audience.

Many marketing researchers use conjoint analysis to identify product characteristics that influence consumers' decisions to purchase certain brands. Trying to uncover the important factors impacting customers' preferences and attribute trade-offs is useful because it allows managers to act on those factors to introduce products and services that meet consumer needs appropriately. For example, a producer of cleaning products may want to know if customers prefer concentrated or diluted liquids, small or large bottles and unique or multi-purpose products. Conjoint analysis (hereafter, CA) results may reveal that consumers prefer a multi-purpose product in concentrated form packed in a small bottle, which allows managers to identify, evaluate and market superior product specifications, increasing likelihoods of success (see Louviere et al. 1990 for a new product application). Traditional CA involves the decomposition of products or services into more elemental aspects called “attributes”, each of which are described by a discrete number of levels. CA researchers systematically manipulate the attribute levels to create attribute level combinations, and each combination describes a product or service (also termed “treatment combinations” or “product profiles”). The resulting combinations are evaluated by consumers (typically, consumers rate or rank the profiles), and their numerical responses allow analysts to quantify the attribute trade-offs (in choice experiments consumers choose their most preferred option). Thus, CA models are “decompositional” in the sense that respondents provide a holistic evaluation or score for each profile, and the analysis of these evaluations reveals the values that individuals associate with each attribute level. Support for the decomposition of preferences can be drawn from Lancaster's view (1966) that preferences for products can be decomposed into separate preferences for each of its characteristics.

Many disciplines (e.g., transport, economics, management, etc) also use CA, which is a “stated preference” (SP) approach. CA’s popularity stems from its advantages relative to observations about consumer choices in real markets, referred to as “revealed preference” (RP) data. The term "revealed preference” refers to preference revelations associated with real market behaviour, while "stated preference” refers to preference revelations associated with stated intentions in hypothetical contexts. As noted by Ben-Akiva, Morikawa and Shiroishi (1991), the advantages of SP over RP are the avoidance of multi-collinearity problems, the possibility of extending the range of attribute levels relative to existing characteristics and the feasibility of testing prototypes or changes to existing products. Other advantages underlined by Hensher, Louviere and Swait (1999) are potential gains in data collection time and expenses, measurement of preferences along dimensions where RP data do not exhibit sufficient variation and possible reductions in bias associated with the failure of RP data to satisfy model assumptions. Lerman and Louviere (1978) demonstrated that there was an empirical link between RP and SP data and models. That is, because SP data are based on consumer's stated intentions, this naturally gives rise to issue of external validity; fortunately, however, a large number of studies have found that SP data mimic choices in real situations (e.g., Ben-Akiva & Morikawa 1990, Hensher & Bradley 1993, Louviere 1993 and Brownstone & Train 1999). Although preferences revealed by SP data were identical to RP data preference revelations in these studies, location and scale calibration were needed. Therefore, many researchers now combine SP and RP data to obtain more accurate estimates of preferences (see Hensher, Louviere & Swait 1999 for more details).

The principal steps in the conduct of CA projects are the following: 1) Define the attributes relevant to describe each option, and their levels (and the competing alternatives in the case of choice experiments). 2) Select the method of elicitation: rating or ranking of alternatives or choice of preferred options. 3) Construct an experimental design in accordance with the two previous points. 4) Decide on the means of administration for the experiment. 5) After collecting the responses, analyse the data with an appropriate model (e.g., analysis of variance, OLS, multinomial logit or probit). The remainder of this chapter discusses each step in more details.

2.1.1 Identification of attributes and levels

The identification of attributes to incorporate in the experiment is an important step because failure to do it carefully can lead to unrealistic and/or incomplete (lack of information) choice situations. Researchers should ensure that all the attributes relevant to decision-making are included in the study without incorporating too many that are unnecessary, since the objective is also to keep the size of surveys manageable. It is worth noting that some attributes not of interest can be omitted from the experiment as long as respondents are informed that all omitted characteristics are identical across products (Yeoh, Uldry, Louviere and Burke 1998). Nevertheless, the story has to be believable enough to ensure that respondents do not make inferences about missing attributes. The selection of levels for each of these identified attributes is also vital to a well-designed experiment. There should be adequate variations in the levels to detect possible effects but the range of values should remain realistic (or at least in accord with the story line). Levels currently faced by consumers should preferably be part of this range to construct believable choice situations. The problem under investigation may dictate the selection of levels; for example, in the case of a new product introduction, the characteristics of prototypes and existing competing products may be included in the experiment. Nevertheless, one should keep in mind that the number of levels partly determines the size of the experiment; and if non-linearity is suspected a minimum of three levels are required. In the case of qualitative attributes, it is important to use clear wording and provide definitions to avoid confusion if necessary. Exploratory research is typically used to complete the attribute and level selection stage; methods of attribute and level identification are discussed by Green and Srinivasan (1978, 1990) and Louviere (1988).

2.1.2 Elicitation procedures

Many elicitation procedures have been used in the past: ratings (e.g., Louviere 1979; Louviere and Meyer 1981), rankings (e.g., Green and Wind 1973; Ben-Akiva, Morikawa and Shiroishi 1991), choice of most preferred option (e.g., Louviere and Woodworth 1983; Hensher and Bradley 1993), and choice of considered alternatives (e.g., Swait and Louviere 1993; Louviere, Fox and Moore 1993), to cite a few. Although ratings and rankings have dominated both academic and commercial applications, more recently discrete or qualitative responses have been applied.

It should be noted that measurement levels differ substantially between elicitation procedures. Category ratings scales are at one extreme in that they are assumed to satisfy at least interval measurement properties for researchers to analyse responses (ratings) with linear models; that is the difference between a 4 and a 5 rating score means the same as the difference between an 8 and a 9. To avoid making this assumption, one may analyse responses as ordinal measures (rather than cardinal), using ordered logit or probit models. However, the use of ratings scales for aggregate analysis further assumes that different individuals interpret rating scales in the same ways, such that a rating of 4 by Zoe means the same as a rating of 4 by Maria. Thus, an aggregate analysis may not satisfy the assumption implied by ratings scales, even if ordered logit and probit are used. Finally, there are questions as to whether individuals can report degrees of preferences accurately, as this is a strong assumption about human cognitive ability. The latter consideration raises questions about the reliability of ratings scales, especially since consumers do not appear to rate products and services in real situations.

Some researchers report that consumers find ranking tasks easier than rating tasks (e.g., Green and Srinivasan 1978). Rankings have the advantage of only assuming ordinal measurement; that is, individuals are able to state which options are preferred to the others but not by how much. Hence, use of rankings allows weaker measurement level assumptions but still offers some depth of information on preferences. However, the reliability of rankings has been questioned by several scholars, such as Ben-Akiva, Morikawa and Shiroishi (1991), who found that responses from different ranking depth were not equally reliable. Beggs, Cardell and Hausman (1981) and Chapman and Staelin (1982) proposed a utility model to analyse ranking data. For example, three soft drinks may be ranked as follows by a respondent: Fanta $\succ$ (preferred to) Coke $\succ$ Sprite. The method of exploded ranks transforms this ranking into two choice sets: (1) Fanta $\succ$ Coke and Sprite, and (2) Coke $\succ$ Sprite; so that from (1) we understand that Fanta is the most preferred drink and from (2) that Coke is preferred to Sprite. However, Chapman and Staelin (1982) found that an explosion depth of 2, which is equivalent to asking respondents to state their most preferred and second best options, appeared relatively safe but an explosion depth of 3 was questionable in their data set. This result appears to support Ben-Akiva et al.'s (1991) findings.

At the other extreme of measurement level lie discrete or qualitative responses such as the most preferred option and considered options. This form of elicitation procedure requires minimal assumptions about respondents' cognitive abilities to report their preferences; moreover, consumers perform this task each time they purchase a product or service and should therefore be familiar with it. Thus, an advantage of choice responses is the ease with which they can be used to forecast market shares and simulate choices, which unlike traditional CA do not require ad hoc transformations of model predictions into choices (e.g., with ratings tasks).

In "consideration" tasks respondents select all the options they would seriously consider from a finite set presented. Such response data may be very beneficial when used in conjunction with a "most preferred" choice task because it adds depth of information without straining respondents' cognitive burden. To refer to the soft drink example, suppose that instead of three options there are five soft drinks in the choice set (Fanta, Coke, Sprite, Dr. Pepper and 7-Up), and respondents are asked to indicate both considered and "most preferred" options. If responses are that Fanta, Coke and Dr. Pepper are considered and Fanta is the most preferred soft drink, the information provided is that Fanta is preferred to all the others, and Coke and Dr. Pepper are preferred to Sprite and 7-Up. In this choice set, rank-ordering the five soft drinks would probably have been more cognitively demanding. At the stage of data cleaning, consideration data also allow researchers to check response validity; that is, if respondents state that their most preferred option is one that they did not consider, there would be grounds for doubting that they took the task seriously. An application using both choice and consideration elicitation procedures is presented in Chapter5.

2.1.3 Experimental design

Statistical design theory is used to manipulate attribute levels and combine them to construct profiles (treatment combinations) or even complete choice sets. Factorial experiments are typically used in marketing, and are constructed to estimate the effect of each attribute on the choice response as independently as possible of the other attributes. In the case of linear regression this objective requires attributes (design matrix columns) to be orthogonal to each other. Let us continue the earlier soft drink example: suppose the drinks are described by three attributes, each of which can take on two values (levels) as shown in Table 2.1.

Table 2.1: Three attributes each with 2 levels describing soft drinks

Attribute	**Level 0**	**Level 1**
Packaging	Can	Plastic bottle
Fridge/Shelf	Fridge	Shelf
Price	$1.50	$2.20

A 2^3 full factorial design can be used to construct all possible profiles with these attributes and levels, as shown in Table 2.2.

Table 2.2: 2^3 full factorial design

	A	**B**	**C**
1	0	0	0
2	0	0	1
3	0	1	0
4	0	1	1
5	1	0	0
6	1	0	1
7	1	1	0
8	1	1	1

In this 2^3 design, columns A, B and C are orthogonal to each other. The design is translated into presentable profiles by taking each row and replacing the 0's and 1's with unique and meaningful levels. Using the attributes and levels in Table 2.1, the first row of the 2^3 design becomes the profile presented in Table 2.3.

Table 2.3: Profile example

Profile 1
Can
Fridge
$1.50

A complete factorial (or full profile) experiment allows one to estimate the attribute main effects and all their interactions independently (in linear regression models), which may be useful to discover the most appropriate model specification. However, as the number of attributes and levels increase, the size of the design may become too large for administration. One way to reduce the number of profiles is to assume that the impacts of higher order interactions are negligible, which may be reasonable in many studies. For example, in most marketing applications, researchers assume that interactions higher than second order provide negligible gains in goodness-of-fit (Green and Srinivasan 1990). Using the eight profiles considered above, we can estimate all main effects and interactions, but fewer profiles are required if one can assume that the 3-way interaction is negligible. Hence, a sample (called a fraction) of profiles from the complete factorial is selected to estimate the effects of interest. If interactions have minor impacts, an experimental design does not need to systematically manipulate them; in other words, columns representing interactions can be assumed constant. We may therefore select a fraction of the complete factorial that does not vary on these dimensions. For example, the complete 2^3 factorial may be reduced by using the 3-way interaction to fractionate the design into two halves of four profiles each; that is, all profiles with this interaction at level 0 are assigned to the first fraction and the others are assigned to the second fraction, as described in Table 2.4. Each fraction is called a fractional factorial design and the 1 in the 2^{3-1} means that it is a half fraction. The last column D is equal to the 3-way interaction ABC and is simply the sum of the columns A, B and C modulo 2 (another way to construct the same matrix would be to start from a 2^{3-2} fractional factorial design and add its foldover). Such a design only permits one to estimate main effects; hence, not only must one assume that interactions are negligible, but they also must be in practice or estimates of the main effects will be biased.

Table 2.4: 2^{3-1} fractional factorial design

	A	B	C	D
1	0	0	0	0
2	0	0	1	1
3	0	1	0	1
4	0	1	1	0
5	1	0	0	1
6	1	0	1	0
7	1	1	0	0
8	1	1	1	1

It is easily observed that each fraction in Table 2.4 does not allow one to estimate the 3-way interaction because it is constant. When sacrificing higher-order interactions to gain degrees of freedom and estimate the effects of interest in fewer choice sets, it is important to realise that those estimated effects may be confounded with higher-order interactions. By "confounded" we mean collinear or correlated; in the 2^{3-1} example, the main effect of A is confounded with the interaction BC, B is confounded with AC and C is confounded with AB. In the soft drink example, if main effects are confounded with two-way interactions, researchers/managers cannot know if an observed effect is due to "packaging" or the "price" and "fridge/shelf" interaction (an issue only if this interaction exists).

In the case of choice experiments, at least two profiles are presented in each choice set; hence, methods have been devised to construct choice sets. A simplistic approach is to randomly assign profiles to choice sets. Another method is to use the "fold-over" to construct the second alternative; that is, the second profile is the image of the first one. In the case of generic (unlabeled) alternatives analysed with MNL models, Bunch, Louviere and Anderson (1996) found that fold-over designs were optimal. In the general case of attributes with M levels, these designs are called shifted designs (also referred to as cyclic designs) in that attributes for additional alternatives are constructed using modular arithmetic applied to the levels of the original attribute column. The number of additional alternatives that may be constructed is a function of the number of attribute levels; more specifically, M attribute levels allows one to construct choice sets of size M. It is also common to design all the alternatives together; design which is referred to as a L^{J*M} design where L is the number of levels, J the number of alternatives and M represents the number of attributes. These designs have the advantage that the attribute columns are orthogonal *within- and between*-alternatives, allowing the estimation of alternative-specific effects and cross-effects (e.g., effect of having Coke at $2.20 on the probability of choosing Fanta). To efficiently estimate the latter effects, Anderson and Wiley (1992) proposed designs with as few as $2M$-1 choice sets. Figure 2.1 presents an example of a choice set from a fold-over design.

Figure 2.1: Choice set example with labeled alternatives

Which drink would you choose? (Tick the appropriate box)

Fanta	Coke
Can	Bottle
Fridge	Shelf
$1.50	$2.20
□	□

It is worth noting that statistical design theory has been developed primarily for linear models (see Atkinson and Donev 1992); however, some approaches have been proposed for designing paired comparison experiments (e.g., El-Helbawy and Bradley 1978; Street, Bunch and Moore 1999), and for designing multiple choice experiments based on MNL models with generic coefficients (Bunch, Louviere and Anderson 1996; Huber and Zwerina 1996). Nevertheless, a lot of research is still needed in the area of experimental design for discrete choice models, especially for situations in which non-IID error components are assumed.

A final point to conclude this section is the inclusion of a "None" alternative (no-purchase option) or an alternative that is constant across all choice sets like "your current brand". The inclusion of the latter may be appropriate when the objective of the research is to identify new product characteristics or changes to existing products that might induce consumers to switch from their current brand to another. Motivations for including "None" options are increased realism in choice tasks and more accurate measures of market shares. However, in some situations, such as choice of transport mode for the daily commute to work, a "None" alternative may not make sense.

2.1.4 Administration Methods

Two points need to be discussed regarding administration methods: (1) information support and (2) methods of administration. Most choice experiments are administered via paper surveys because this is accessible to virtually everyone and does not require respondents to use special facilities. In this case, choice sets typically are described as shown in Figure 2.1. More recently, new technology allows presentation of choice sets on computer screens; obviously the feasibility of this approach depends on the sample population targeted because not all consumers are computer literate. Other experiments display choice set information in the form of physical products; for example, this may be especially useful if product shape or colour are attributes of interest. Audio and video information also can be used but costs and time constraints may limit such applications. Administration of surveys can be by mail, email, Internet or personal interviews, depending on the problem at hand and the type of information that needs to be displayed. It is worth noting that electronic administration allows one to check response inconsistencies directly and prompt respondents to reconsider invalid responses, which should increase data quality.

2.1.5 Data analysis

The choice of models used for data analysis depends mostly on decisions made in earlier stages of the research such as the choice of elicitation procedures. Ratings are typically analysed with linear regression models because many researchers assume that ratings have interval measurement properties. However, ordered logit or probit can also be used to analyse and model rating responses if they are treated as ordinal rather than cardinal, and rankings can be exploded and analysed with MNL models. The discussion that follows concentrates on MNL and MNP models because the research in this thesis deals with *choice* experiments.

The MNL model was first used with binary discrete outcomes because it was a simple way to ensure that a model would predict in the 0 to 1 range. Its main advantage is the simplicity with which it can be estimated due to having a closed form expression for the choice probabilities; hence, estimation software for the MNL model is included in many statistical packages. In marketing, the MNL model is attractive because one can compute the probabilities of specific random utility models.

Random Utility Models, or more generally Random Utility Theory (RUT), was introduced by Thurstone (1927) for paired comparisons and extended by McFadden (1973) to multiple comparisons (choices). In the RUT framework consumers are assumed to associate a certain level of utility (U_{ij}) with each alternative in a choice set and choose the most preferred alternative as if they are rational in the sense that they choose the alternative that maximises their utility. It is typically assumed that utility can be decomposed into linear functions of attributes (x_{ij}) that describe alternatives. These assumptions lead to the following expression:

$$U_{ij} = \alpha_j + x_{ij}\beta_j + \varepsilon_{ij} \qquad (2.1)$$

where U_{ij} is an unobserved indirect utility for individual i and alternative j, $j = 1,..J$,

α_j is a vector of alternative-specific constants (ASC's)

x_{ij} is a vector of independent variables for individual i and alternative j,

β_j is a vector of parameters,

ε_{ij} is an error component.

The indirect utilities U_{ij} for individual i and alternative j ($j = 1,2,...,J$) are unobserved variables because only discrete choices are reported. β_j represents consumer trade-offs among different attributes of a specific product, and captures what is known as the systematic component of utility. Finally, an error component ε_{ij} is introduced to reflect that analysts do not know consumer preferences with certainty. This randomness may occur because of unobserved attributes or taste variations, measurement errors, imperfect information and/or proxy variables.

Although the U_{ij}'s are unobserved, RUT assumes that consumers choose the alternative they prefer most, as follows:

$U^*_{ij} = 1$ if $U_{ij} > U_{ik}$ for any $j \neq k$

$U^*_{ij} = 0$ otherwise.

U^*_{ij} represents the choice indicator.

McFadden (1973) showed that if error components are IID Gumbel (extreme value type I) distributed, that is

$$f(\varepsilon) = \mu\, e^{-\mu(\varepsilon-\eta)} \exp\left[-e^{-\mu(\varepsilon-\eta)}\right],$$

where μ is a positive scale parameter and η is a location parameter, the choice probabilities are identical to the MNL choice probabilities. Some properties of the Gumbel distribution are that its mode is η, its mean is $\eta + \gamma / \mu$ (where γ is Euler's constant) and its variance is $\pi^2/6\mu^2$. Thus, the probability of individual i choosing alternative k is expressed as in equation (2.2) (see Ben-Akiva and Lerman 1985).

$$P\left(U_{ik} > U_{ij}\ \forall\ j \neq k\right) = \frac{\exp \mu\left(\alpha_{ik} + x_{ik}\beta_j\right)}{\sum_{j=1}^{J} \exp \mu\left(\alpha_{ij} + x_{ij}\beta_j\right)} \tag{2.2}$$

The MNL model is a special case of the Luce model (Luce 1959), in which the utility function is additive in the attributes x_{ij}.

Because the utilities U_{ij} are latent, two identification issues arise: the first is due to the fact that the addition of a constant to all utilities does not change observed choices. For example, suppose that three soft drinks are presented in a choice set, Fanta, Coke and Sprite; suppose that the unobserved utilities are 20, 16 and 10 for the three drinks respectively. However, the researcher can only observe that Fanta was chosen, and if 5 points are added to each utility, the resulting utilities are perfectly consistent with the observed choice. Therefore, only differences in utility can be estimated. Typically, the utility of one of the options is set as the base and the others become differences from this base. Hence, one ASC α_j is set equal to zero. The second identification issue comes from the fact that the multiplication of all three utilities by a positive constant does not change their ordering, again leaving the observed responses unchanged. Thus, one model parameter has to be fixed to overcome the identification issues; fixing one coefficient would solve the problem, but if it has the wrong sign, all other coefficients also will have wrong signs. So, the variance typically is fixed by setting $\mu = 1$.

As noted earlier, MNL models are restrictive because they assume independently and identically distributed (IID) error components; in the case of MNL, this leads to the Independence of Irrelevant Alternative (IIA) property. The latter implies that the odds of choosing one option over a second do not depend on any remaining options. For example, if the odds of choosing Coke relative to Fanta (probability of choosing Coke divided by the probability of choosing Fanta) equals 0.6 in a three option choice set, but equals 0.4 if Pepsi is added to the set, the IIA property would be violated. The IIA assumption has been found to be too restrictive in some applications (Huber, Payne and Puto 1982; Simonson and Tversky 1992), although Debreu (1960) had previously proved that Luce models (hence, also MNL models) over-predict the joint probability of selection of two options that are perceived to be similar rather than independent. Models that are not affected by the IIA property include Nested logit and MNP. The latter has a very general specification form, which favours its use in our research.

If one assumes that the error components ε_{ij} in equation (2.1) are distributed as normal random variates, the resulting model is called multinomial probit (MNP), which was first proposed by Thurstone (1927). The identification of the MNP model is typically achieved by setting the first element of its error covariance matrix equal to 1. MNP models allow one to relax the IID assumption on error components but require the evaluation of integrals of dimension ($J-1$); thus, when $J>4$, the probabilities cannot be evaluated numerically, and one must use simulation methods (e.g., Keane 1992). One very successful simulator is the GHK simulator (after Geweke, Hajivassiliou and Keane), which Hajivassilou, McFadden and Ruud (1996) found to be quick and reliable. To avoid the problem of high-dimension integrals implied by Maximum Likelihood estimation, Bayesian methods also may be used. Bayesian methods are used in the remainder of this thesis and introduced in the next section.

2.2 Bayesian Estimation Methods

This section introduces Bayesian estimation methods. A generic discussion of Bayesian analysis is not easy because the steps that one undertakes are contingent on the problem at hand. However, what can be said is that the objective of much Bayesian analysis is to obtain posterior distributions of the parameters of interest. So, I introduce the concept by means of an example that shows how to obtain analytically the relevant posterior distributions in a regression analysis. Section 2.3.2 discusses and demonstrates the use of the Gibbs sampler in Bayesian analysis and the reason why this sampler revived Bayesian methods.

Prior and posterior distributions, as well as marginal and joint distributions are used extensively in Bayesian methods. A quick explanation is therefore in order.

Table 2.5 presents an example of the marginal and joint distributions of two binary variables x and y.

Table 2.5: Marginal and joint probabilities of x and y (example)

	$x=0$	$x=1$	Marginal probability of y
$y=0$	$p_{x=0,y=0}=0.3$	$p_{x=1,y=0}=0.4$	$p_{y=0}=0.7$
$y=1$	$p_{x=0,y=1}=0.1$	$p_{x=1,y=1}=0.2$	$p_{y=1}=0.3$
Marginal probability of x	$p_{x=0}=0.4$	$p_{x=1}=0.6$	1

In table 2.5 the joint probabilities of x and y are denoted by p_{xy} and the marginal probabilities by p_x and p_y respectively. The marginal distribution p_x is simply the sum of the joint probabilities over all values of y for each value of x; thus, $p_{x=0} = p_{x=0,y=0} + p_{x=0,y=1}$. Similar calculations are done for the marginal distribution of y. In this example the variables of interest are discrete (0 or 1). With continuous variables the computations are similar but the summation over all values of a continuous variable requires an integral to be evaluated. Therefore, if x and y are continuous, the marginal distribution is computed by evaluating $p_x = \int p_{xy} dy$ over the domain of definition of y.

As noted earlier, Bayesian analysis is concerned with the marginal posterior distributions of some parameters of interest. As its name suggests, the prior distribution of a variable corresponds to a priori knowledge about the variable in question. For example, in marketing a price coefficient can be assumed to be negative a priori, and this knowledge can be used and accounted for in the analysis of the data. The posterior mean of the price coefficient is computed taking two sources of information: 1) the data and 2) the prior information. Distributions have both means and variances, so one can also include prior information on the variance. For example, a tight prior variance means that the researcher has a very precise idea (a priori) of the price coefficient, and the data has less impact on the posterior mean. On the contrary, a very large prior variance reflects a lack of a priori knowledge and the posterior mean mostly relies on the data. Completely flat priors represent a complete lack of prior information, and have therefore no bearing on the posterior mean. The next section uses these preliminary concepts to perform a Bayesian regression analysis.

2.2.1 Bayesian Regression

Univariate and multivariate regressions have been studied extensively. Frequentist methods such as Maximum Likelihood and Ordinary Least Squares have been applied to solve this problem, and their respective solutions are well-known; both are almost identical. Not surprisingly, Bayesian analysis produces a similar solution but it is interesting to follow the step-by-step process needed to obtain such solutions because it nicely illustrates the fundamentals of the Bayesian machinery.

Consider the following model that might represent a marketing application, such as relating sales (y) to advertising expenditures and shelf space (X):

$$y = X\beta + \varepsilon$$

where y is a vector containing n observations y_i (such as sales of each store)

X is a matrix containing the explanatory variables

ε is vector of normally distributed errors with zero mean and variance σ^2.

This implies that y_i is normally distributed with mean $X_i\beta$ and variance σ^2. Therefore, the distribution of y conditional on β and σ^2 is

$$p\left(y \mid \beta, \sigma^2\right) = \left(2\pi\sigma^2\right)^{\frac{-n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\varepsilon'\varepsilon\right\}.$$

The objective now is to work out the marginal distribution of both β and σ^2. Using Bayes rule, namely $p(a|b)p(b) = p(b|a)p(a) = p(a,b)$ the following two conditional distributions can be derived:

$$p(\beta | y) = \frac{p(y|\beta)p(\beta)}{p(y)} \tag{2.3}$$

$$p(\sigma^2 | y) = \frac{p(y|\sigma^2)p(\sigma^2)}{p(y)} \tag{2.4}$$

At this stage, one may wonder what $p(\beta)$ and $p(\sigma^2)$ represent. They are referred to as "priors", and can be thought of as a priori information that the researcher may have about β and σ^2 coming from prior research, theory or belief. However, in order to compare the results with the classical way of approaching this problem (which does not involve prior distributions), the priors are assumed to be uninformative (also called "flat" priors). In other words, it is assumed that the researcher simply has no information on the parameters a priori. The priors in this example are assumed to be proportional to a constant and therefore are not proper densities. Note that $p(\sigma^2)$ is made non-informative on $\log(\sigma^2)$. In more complex models, priors cannot always be assumed to be improper, but a proper density can be imposed with a large prior variance, which in effect makes prior knowledge very diffuse. Such priors also are called "diffuse" priors.

Suppose that:

$p(\beta) \propto cst$ (constant)

$$p\left(\sigma^2\right) \propto \frac{1}{\sigma^2}.$$

It is important to realise that, although conditional distributions such as $p(\beta \mid y)$ and $p\left(\sigma^2 \mid y\right)$ need to be determined and evaluated, all constants of proportionality can be dropped, hence the symbol "proportional to" ($\propto$) is used to simplify the exposition. For instance, $p(\beta \mid y)$ is a function of $p(y)$ and $p(\beta)$ which are both constant. $p(\beta)$ is constant because it represents a flat prior and $p(y)$ represent data which are a fixed quantity.

Equations (2.3) and (2.4) can therefore be expressed as follows:

$$p(\beta \mid y) \propto p(y \mid \beta)$$

$$p\left(\sigma^2 \mid y\right) \propto p\left(y \mid \sigma^2\right).$$

Using Bayes rule and assuming uninformative priors, it is thus straightforward to write the marginal parameter distributions.

The next stage is critical since $p(\beta \mid y)$ needs to be evaluated while knowing only $p\left(y \mid \beta, \sigma^2\right)$. $p(\beta \mid y)$ is recognised to be proportional to $p(y \mid \beta)$, which remains unknown. The following relationship can be used to evaluate $p(y \mid \beta)$ from $p\left(y \mid \beta, \sigma^2\right)$:

$$p(y \mid \beta) = \int p\left(y \mid \beta, \sigma^2\right) d\sigma^2$$

$$\int p\left(y \mid \beta,\sigma^2\right) d\sigma^2 = \left(2\pi\right)^{\frac{-n}{2}} \int \left(\sigma^2\right)^{\frac{-n}{2}} \exp\left\{-\frac{\varepsilon'\varepsilon}{2\sigma^2}\right\} d\sigma^2 \tag{2.5}$$

The integral is inverse gamma in σ^2 with parameters $a=\frac{n}{2}$, $b=\frac{\varepsilon'\varepsilon}{2}$. This observation is useful because integrating over the whole domain of a distribution function simply equals 1. Equation (2.5) can therefore be expressed as follows:

$$p\left(\beta \mid y\right) \propto \Gamma\left(a\right) b^{-a} \tag{2.6}$$

The error sum of squares $\varepsilon'\varepsilon$ is quadratic in β. It is therefore possible to find λ, $\hat{\beta}$ and c such that:

$$\left(\beta-\hat{\beta}\right)' \lambda\left(\beta-\hat{\beta}\right)+c=\varepsilon'\varepsilon=\left(y-X\beta\right)'\left(y-X\beta\right)=y'y-2\beta'X'y+\beta'X'X\beta .$$

Expanding the left-hand side yields: $\beta'\lambda\beta-2\beta'\lambda\hat{\beta}+\hat{\beta}'\lambda\hat{\beta}+c$

The comparison of both expansions leads to $\lambda=X'X$ and $\lambda\hat{\beta}=X'y$, which imply that $\hat{\beta}=\left(X'X\right)^{-1}X'y$ and $c=y'y-\hat{\beta}'X'X\hat{\beta}$. Using this result in (2.6) and dropping $\Gamma\left(a\right)$ since it is constant we can write:

$$p\left(\beta \mid y\right) \propto \left(1+\frac{\left(\beta-\hat{\beta}\right)' X'X\left(\beta-\hat{\beta}\right)}{y'y-\hat{\beta}'X'X\hat{\beta}}\right)^{\frac{-n}{2}},$$

which is multivariate t-distributed (Bernardo and Smith 1994, page 435).

The second distribution $p\left(\sigma^2 \mid y\right)$ can be obtained in the same manner.

$$p\left(y \mid \sigma^2\right)=\int p\left(y \mid \beta,\sigma^2\right)\frac{1}{\sigma^2}\,d\beta$$

$$\int p\left(y \mid \beta,\sigma^2\right) d\beta = \left(2\pi\sigma^2\right)^{\frac{-n}{2}-1} \exp\left(\frac{-1}{2\sigma^2}c\right) \int \exp\left(\frac{-1}{2\sigma^2}\left(\beta-\hat{\beta}\right)' X'X\left(\beta-\hat{\beta}\right)\right) d\beta$$

The integral is normal in β ; hence, the multiplication of the integral by the appropriate constant of proportionality makes it collapse to unity. The constant of proportionality is simply $\left(2\pi\right)^{\frac{-q}{2}}\left|X'X/\sigma^2\right|^{\frac{-1}{2}}$, where q represents the dimension of β and therefore,

$$p\left(\sigma^2 \mid y\right) \propto \left(2\pi\sigma^2\right)^{\frac{-n}{2}-1} \exp\left(-\frac{1}{2\sigma^2}\left(y'y-\hat{\beta}'X'X\hat{\beta}\right)\right)\left(2\pi\sigma^2\right)^{\frac{q}{2}}\left|X'X\right|^{\frac{1}{2}}$$

$$p\left(\sigma^2 \mid y\right) \propto \left(2\pi\sigma^2\right)^{-\left(\frac{n-q}{2}+1\right)} \exp\left(-\frac{1}{2\sigma^2}\left(y'y-\hat{\beta}'X'X\hat{\beta}\right)\right),$$

which is inverse gamma with parameters $\frac{n-q}{2}$ and $\frac{1}{2}\left(y'y-\hat{\beta}'X'X\hat{\beta}\right)$. This concludes the demonstration of using Bayesian analysis to solve a regression problem.

Although quite straightforward in the case of regression, working out the marginal distribution of each parameter can quickly become burdensome or impossible in more complex problems. Here, in order to get the marginal distributions of β and σ^2 only two integrals need to be evaluated. More parameters introduce more integrals, some of which may be unknown. This typically occurs in more complex models such as MNL and MNP considered in chapters 3, 4 and 5. However, a new "tool", namely the Gibbs sampler, revolutionised Bayesian analysis by allowing one to generate draws from marginal distributions (i.e. $p\left(\beta \mid y\right)$ and $p\left(\sigma^2 \mid y\right)$ above) without the need to analytically solve the integrals. The Gibbs sampler is introduced in the next section.

2.2.2 The Gibbs Sampler

A crude description of the Gibbs sampler is that it is a method for generating draws from a marginal distribution without having to calculate the density explicitly. This section only describes how to implement the Gibbs sampler without going into the mathematical details of why it actually works. A very good explanation can be found in Casella and George (1992).

According to Casella and George (1992) the first discussion of the Gibbs sampler can be found in Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953). However, a more recent exposition of the algorithm by Gelfand and Smith (1990) motivated new research using the Gibbs sampler.

The previous section showed how to obtain $p(\beta \mid y)$ and $p(\sigma^2 \mid y)$ for the case of linear regression. Both distributions can be computed by integrating $p(y \mid \beta, \sigma^2)$ over β and σ^2 respectively. The Gibbs sampler allows one to draw from $p(\beta \mid y)$ and $p(\sigma^2 \mid y)$ without having to work out the integrals. So, given $p(\beta \mid y, \sigma^2)$ and $p(\sigma^2 \mid y, \beta)$, the method allows one to directly draw from $p(\beta \mid y)$ and $p(\sigma^2 \mid y)$. Specifically, the sampler shows how to draw from $p(\beta \mid y, \sigma^2)$ and $p(\sigma^2 \mid y, \beta)$ to obtain a draw from the joint posterior distribution $p(\beta, \sigma^2 \mid y)$. The mechanics of the methods are described below for the regression analysis case reported in the previous section. Assuming uninformative priors, recall from the previous section that:

$$p(\beta \mid \sigma^2, y) \propto p(y \mid \beta, \sigma^2) \propto (\sigma^2)^{\frac{-n}{2}} \exp\left\{-\frac{\varepsilon'\varepsilon}{2\sigma^2}\right\} \propto \exp\left\{-\frac{\left[(\beta-\hat{\beta})' X'X(\beta-\hat{\beta})\right]}{2\sigma^2}\right\}$$

$$p(\sigma^2 \mid \beta, y) \propto p(y \mid \beta, \sigma^2) \propto (\sigma^2)^{\frac{-n}{2}} \exp\left\{-\frac{\varepsilon'\varepsilon}{2\sigma^2}\right\}$$

Therefore $\beta \mid \sigma^2, y$ is normally distributed $N\left(\hat{\beta}, (\sigma^2 XX)^{-1}\right)$ and $\sigma^2 \mid \beta, y$ is inverse gamma distributed $IG\left(\frac{n}{2}, \frac{\varepsilon'\varepsilon}{2}\right)$.

The Gibbs sampler simply requires one draw from each conditional distribution in turn, as follows.

1. Define starting values $\beta^{[0]}$ and $\sigma^{2[0]}$ for β and σ^2. Following the example, we may define $\beta^{[0]} = 0$ and $\sigma^{2[0]} = 1$. Further define $i = 1$ and

2.a. generate $\beta^{[i]}$ from $N\left(\hat{\beta}, \left(\sigma^{2[i-1]} XX\right)^{-1}\right)$

2.b generate $\sigma^{2[i]}$ from $IG\left(\frac{n}{2}, \frac{\varepsilon'\varepsilon}{2}\right)$

3. $i = i + 1$, go back to 2.a.

Steps 2 and 3 are repeated twice, N_1 and N_2 times. After the first N_1 iterations the sampler can be assumed to have converged to the joint posterior distribution. In the second stage, draws can be assumed to come directly from the posterior and can therefore be collected for inference. The choice of N_1 may seem controversial since, in theory, this sampling scheme converges to the joint posterior as the number of draws tends to infinity. However, in practice it is straightforward to monitor the convergence and it is typically observed that the chain is drawing from the same region after a few hundred draws. In models such as the ones developed in the next chapters of this manuscript, it is common to choose $N_1 = N_2 = 10,000$, even though the chain appears to have converged in much fewer draws.

In more complex problems, the choice of starting values may be important. Specifically, it is not always possible to monitor convergence of an iterative simulation from a single sequence. The major difficulty, known as slow convergence, is that for many iterations the random walk can remain in a region heavily influenced by the starting distribution. Gelman and Rubin (1992) and Gelman (1995) argue that multiple starting values should always be considered to better assess convergence. They propose the creation of an overdispersed estimate of the target distribution from which multiple starting values are drawn and used to get multiple sequences of the iterative simulation. This approach allows the construction of an estimate of how much more accurate the results may become if the number of iterations is increased. Furthermore, considering multiple sequences can reveal the variability due to the starting values as well as the variability of the estimates themselves.

Chapter 3

Sampling the Covariance Matrix in the Multinomial Probit Model

Abstract

This chapter presents a new method for carrying out Bayesian inference for a multinomial probit model, which permits a parsimonious representation of the inverse of the covariance matrix. The inverse of the covariance matrix is decomposed using a Cholesky factorization and element selection is performed on the lower triangle of the Cholesky factor. All parameter estimates are obtained by model averaging, with the computation carried out using a Markov chain Monte Carlo sampling scheme which requires draws from the normal and gamma distributions only. The methodology is illustrated with an example.

3.1 Introduction

The multinomial probit (MNP) model has been investigated by a number of researchers. Parameter identification was discussed by Bunch (1991) and Keane (1992) and the model has been estimated by maximum likelihood and by Bayesian methods. However, there are still unresolved issues in estimating the MNP model. For example, Keane (1992) argues that even if the model is formally identified, it is still necessary to understand the conditions under which the estimates are well behaved in the sense that the likelihood is not flat near its maximum. Current methods for maximizing the likelihood of the MNP model are computationally expensive because it is necessary to evaluate the choice probabilities numerically. This makes evaluation of the choice probabilities using numerical integration infeasible in high dimensions. Simulation methods such as those of Geweke-Keane-Hajivassiliou (see Geweke 1991; Hajivassiliou and Ruud 1994; Keane 1994) evaluate these probabilities accurately, but are demanding computationally.

Bayesian methods using Markov chain Monte Carlo (MCMC) with data augmentation (Albert and Chib 1993; McCulloch and Rossi 1994) are an appealing option because it is unnecessary to evaluate probabilities. The main difficulty with most MCMC methods is in sampling from the covariance matrix because the covariance matrix is usually constrained to identify the parameters. Chib, Greenberg and Chen (1998) show how to use the Metropolis-Hastings (MH) algorithm to generate the covariance matrix under any prior. Our paper proposes and illustrates a straightforward way to sample the covariance matrix that is more flexible than previous approaches, but at the same time is very fast. Our method is based on a Cholesky decomposition of the inverse of the covariance matrix and allows a parsimonious representation of the inverse by using element selection on the off-diagonal elements of the Cholesky factor. All the regression parameters and the elements of the covariance matrix are estimated by model averaging with the computation carried out by Markov chain Monte Carlo.

The rest of the paper is organized as follows. Section 2 outlines the problem of discrete choice modeling, Section 3 presents our method without element selection and in the differenced space. We do so to relate the results obtained using our approach to those in Chib, et. al. (1998). Section 4 illustrates the method in Section 3 with a travel choice example. Section 5 introduces the element selection method and a parameterization of the MNP model in the original undifferenced space and applies it to the same data.

3.2 Methodological Background

A multinomial response (or choice) is defined as a choice from a set of options; i.e., subjects choose exactly one option from a set of options (each set is called a "choice set"). The choice outcomes constitute a discrete multinomial dependent variable, and are explained by covariates that characterise the options, the individual choice makers themselves, or both. That is, suppose there are J choices, and define $y_{ij}=1$ if the ith individual chooses option j, and let $y_{ij}=0$ otherwise. We define a multinomial response model as follows:

$$U_{ij}^{*}=x_{ij}^{*}{}'\beta_{j}^{*}+\varepsilon_{ij}^{*} \tag{3.1}$$

where U_{ij}^{*} is an unobserved variable for individual i and option j, $j=1,2,\ldots,J$,

x_{ij}^{*} is a vector of independent variables for individual i and option j,

β_{j}^{*} is a vector of parameters,

ε_{ij}^{*} is an error term.

The U_{ij}^{*}'s can be thought of as indirect utilities associated with the options. They are unobserved, but according to Random Utility Theory

$y_{ij}=1$ if $U_{ij}^{*}>U_{ik}^{*}$ for any $j\neq k$,

$y_{ij}=0$ otherwise.

McFadden (1973) demonstrated that if the errors are distributed as Gumbel random variates, then the parameters can be estimated with a Multinomial Logit (MNL) model. Multivariate normal random errors lead to the Multinomial Probit (MNP) model. The

MNL model is widely used because the choice probabilities have a closed form; and hence it is straightforward to obtain maximum likelihood estimates.

Direct evaluation of the likelihood for the MNP model requires multidimensional integration to evaluate the probabilities, and the number of integrals increases with the number of choice options. Several simulation methods have been proposed to evaluate the probabilities for a given value of the parameters because numerical integration becomes infeasible as the number of choice options increases. The best known method is the GHK simulator proposed by Geweke (1991), Hajivassiliou and McFadden (1998) and Keane (1994). The simulated probabilities are unbiased, but the log likelihood is not (Börsch-Supan and Hajivassiliou 1993). This may cause simulated maximum likelihood (SML) estimates obtained by this method to be biased, although the method seems to work well in practice (Geweke, Keane and Runkle 1994).

The MNP model is more difficult to estimate than MNL, but its richer error structure allows relaxation of the assumption of the IID errors that leads to the Independence of Irrelevant Alternatives or (IIA) property of MNL and related models (e.g., Luce 1959). The IIA property states that the ratio of the probabilities of any two options is independent of the presence or absence of other options. However, there are two identification issues in the MNP model, because the dependent variable is not measured on a cardinal scale. That is, only $(J-1)(J-2)/2-1$ elements in the matrix can be estimated (Bunch 1991), which can be understood intuitively by noting that the choice probabilities are invariant to adding a constant to both sides of (3.1) or multiplying each side of (3.1) by a positive constant. Identification is achieved by setting the utility of one option to zero (e.g., the J th one) and restricting one element of the covariance matrix (e.g., element 1,1). This is accomplished, for example, by subtracting U^*_{iJ} from U^*_{ij} to obtain,

$$W_{ij} = U^*_{ij} - U^*_{iJ} = x^*_{ij}{}'\beta^*_j - x^*_{iJ}{}'\beta^*_J + \varepsilon^*_{ij} - \varepsilon^*_{iJ} = x'_{ij}\beta_j + \varepsilon_{ij} \qquad (3.2)$$

which, in vector notation, becomes

$$W_i = X_i\beta + \varepsilon_i$$

where $\varepsilon_i \sim N(0,\Sigma)$, and by setting $\Sigma_{11} = 1$.

As mentioned above, one way to estimate such a structure is to use SML. Bayesian methods using data augmentation and with the computation carried out by MCMC provide a second approach. For example, Albert and Chib (1993) estimated binary probit ($J = 2$) using data augmentation by recognizing that, conditional on β and y_i, W_i has a truncated normal distribution with mean $X_i\beta$ and variance equal to 1. This simplifies the estimation problem because conditional on W_i the problem is Gaussian. In other words, once a value for W_i is drawn, the estimation problem is reduced to a seemingly unrelated regressions problem (SUR). A Markov chain is formed by drawing in turn, values for W_i and β. This approach is easily implemented and works particularly well with the binary probit model because the covariance matrix reduces to a fixed scalar equal to 1. However, if the number of options is larger than 2, and a general error covariance is assumed, then the sampling scheme requires draws from the covariance matrix Σ, whose conditional distribution is complex because of the constraint that $\Sigma_{11} = 1$.

McCulloch and Rossi (1994) solved this problem by drawing unconstrained values of Σ and the regression coefficients β and obtained proper posterior distributions for Σ and β by setting proper priors for them. However, it may not be straightforward to set proper priors for Σ and β because Σ and β are not independent in the prior due to the identification constraints (see McCulloch, Polson and Rossi 2000, section 5.1). McCulloch, Polson and Rossi (2000) use fully identified parameters, but do not have the flexibility of our approach to carry out element selection on a factorization of Σ or Σ^{-1}. Chib, Greenberg and Chen (1998) use a fully identified model and the Metropolis-Hastings algorithm to sample directly from the conditional distribution of Σ. Their approach allows for flexible priors on the identified elements of Σ, but they do not use element selection on a factorization of Σ. Sections 3 and 5 present a new sampling method that allows a flexible prior for Σ, but is simpler to implement than the method of Chib, Greenberg and Chen (1998). As mentioned previously, the method is based on a decomposition of the inverse of the covariance matrix and a selection procedure is implemented on the elements of the decomposition. Introducing a selection approach is motivated by the fragile identification of the model. Specifically the aim of the method is to select and estimate significant elements of the Cholesky decomposition of the inverse of the covariance matrix. In particular, our approach to parsimony becomes attractive when the number of alternatives is large because the number of unknown elements in the covariance matrix increases quadratically with the number of options. For example, if the data appear to support a diagonal structure, then our method would converge to it and set the non-diagonal elements of the covariance matrix to zero.

Variable selection methods were first introduced to determine which independent variables enter into a linear regression model (Mitchell and Beauchamp 1988; George and McCulloch 1993, 1997; Raftery, Madigan and Hoeting 1997). The ideas were developed further by Smith and Kohn (1996) who applied them to estimating regression functions non-parametrically. Our approach is based on that of Smith and Kohn (1999) who estimate a covariance matrix using a Cholesky decomposition of its inverse and apply element selection to the lower triangle of the Cholesky factor. Smith and Kohn (1999) do not, however, consider the multinomial probit model, which has the added complications of identifying constraints on the covariance matrix.

3.3 Sampling from the Posterior Distribution in the MNP Model

This section presents the three steps used to draw from the posterior distribution in the MNP model. Steps 1 and 3 are usually broken down further to make the sampling tractable. The third step introduces the basic representation of the inverse covariance matrix required for our method. In this section and Section 4 we work in the difference space in order to relate our method to previous approaches that also used differencing. An alternative way of estimating the MNP is to use the original space with extra constraints on the covariance matrix. This approach is outlined in Section 5.

Sample W_i from $\pi(W_i \mid y_i, \beta, \Sigma)$,

Sample β from $\pi(\beta \mid W, \Sigma)$,

Sample Σ from $\pi(\Sigma \mid W, \beta)$,

where $W = (W_1, \ldots, W_N)$ and N is the number of choice sets.

Each of these conditional distributions is now considered in more detail. The distribution represented in step one is a truncated multivariate normal distribution. Define $W_{i(-j)}$ as the vector W_{ij} excluding the jth component. The elements W_{ij} can be sampled by drawing in turn from univariate normal distributions $\pi(W_{ij} \mid y_{ij}, W_{i(-j)}, \beta_j, \Sigma)$, truncated to the region implied by the indicator variable y_{ij} (see McCulloch and Rossi 1994 for more details). It is also possible to use an accept-reject method to draw directly from the vector W_i, but if the number of options J is large, this method generates too many draws that are rejected before obtaining one that satisfies the constraints.

The conditional distribution in step 2 is

$$\pi(\beta \mid W, \Sigma) \propto \exp\left(-\frac{1}{2}\sum_i \varepsilon_i' \Sigma^{-1} \varepsilon_i\right) \pi(\beta).$$

The term

$$\exp\left(-\frac{1}{2}\sum_i \varepsilon_i' \Sigma^{-1} \varepsilon_i\right)$$

is Gaussian in β; making it straightforward to generate β. In particular, if the prior density of β is $N\left(\beta_0, V^{-1}\right)$, then $\beta \mid W, \Sigma \sim N\left(\hat{\beta}, \Delta^{-1}\right)$,

where $\Delta = \sum_i X_i' \Sigma^{-1} X_i + V$ and $\hat{\beta} = \Delta^{-1}\left(\sum_i X_i' \Sigma^{-1} W_i + V\beta_0\right)$.

Step three introduces our new approach to sample from the constrained covariance matrix Σ. Our method is based on a decomposition of the inverse of Σ as

$$\Sigma^{-1} = B' D B, \tag{3.3}$$

with D a diagonal matrix and B a lower triangular matrix with ones on the diagonal. To enforce the identification of Σ through the condition $\Sigma_{11}=1$, we assume, without loss of generality, that $D_{11}=1$. The matrix Σ is generated by generating B and D. This decomposition of the inverse covariance matrix is used by Smith and Kohn (1999) in estimating seemingly unrelated regression models, but these authors do not consider the application to the MNP model, which requires further identification conditions.

We follow Smith and Kohn (1999) and first show how to sample from the conditional distribution $\pi(B \mid W, \beta, D)$:

$$\pi(B \mid W, \beta, D) \propto |B'DB|^{N/2} \exp\left(-\frac{1}{2}\sum_i \varepsilon_i' B' D B \varepsilon_i\right) \pi(B \mid D)$$

which simplifies to

$$\pi(B \mid W, \beta, D) \propto \exp\left(-\frac{1}{2} \operatorname{tr} BAB'D\right) \pi(B \mid D), \tag{3.4}$$

because $\det(B)=1$. In (3.4), the matrix $A=\sum_i \varepsilon_i \varepsilon_i'$.

For convenience, we partition the matrices B, D and A as follows:

$$B=\begin{bmatrix} 1 & 0 & 0 & 0 \\ b_2' & 1 & 0 & 0 \\ \vdots & & 1 & 0 \\ b_{J-1}' & & & 1 \end{bmatrix} \quad D=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & d_{J-1} \end{bmatrix} \quad A=\begin{bmatrix} A_i & a_i & . \\ a_i' & \alpha_i & . \\ . & . & . \end{bmatrix}$$

The vector b_i is $(i-1)\times 1$, and A_i and a_i are dimensioned conformally as $(i-1)\times(i-1)$ and $(i-1)\times 1$. Using this notation,

$$\pi(b_1,\dots,b_{J-1} \mid W,D,\beta) \propto \prod_{i=2}^{J-1} \exp\left(-\frac{1}{2} d_i(b_i' A_i b_i + 2b_i a_i)\right) \pi(B \mid D) \qquad (3.5)$$

That is, up to the prior $\pi(B \mid D)$, the vectors $b_1,\dots,b_{J-1}$ are independent and Gaussian; and hence it is easy to generate them irrespective of the form taken by the prior $\pi(B \mid D)$. In particular, if the b_i are normal and independent a priori with $b_i \sim N\left(b_{0,i}, B_{0,i}^{-1}\right)$, then from (3.5), $\pi(b_i \mid W,\beta,D)$ is $N\left(\tilde{b}_i, B_i^{-1}\right)$, with $B_i = A_i d_i + B_{0,i}$ and $\tilde{b}_i = -B_i^{-1}\left(a_i d_i - B_{0,i} b_{0,i}\right)$.

Finally, we show how to generate the diagonal elements $d_2,\dots,d_{J-1}$ of D. The conditional density of these elements is given by

$$\pi(d_2,\dots,d_{J-1} \mid W,\beta,B) \propto \prod_{i=2}^{J-1} d_i^{N/2} \exp(-\frac{1}{2} d_i C_{ii}) \pi(D \mid B),$$

where $C = BAB'$.

That is, up to the prior $\pi(D \mid B)$, the d_i have independent gamma distributions. It is, therefore, straightforward to generate $d_2,\dots,d_{J-1}$ irrespective of the form of the prior $\pi(D \mid B)$. In particular, if the prior density of d_i is $\text{Gamma}(\delta_0, \Delta_0)$ and the d_i are independent then $\pi(d_i \mid W,\beta,B)$ is $\text{Gamma}(\delta, \Delta)$, where $\delta = 0.5N + 1 + \delta_0$ and $\Delta = 0.5C_{ii} + \Delta_0$.

It is now clear that using the representation (3.3) of the covariance matrix, it is straightforward to draw B and D for any prior $\pi(B,D)$ because it is always possible to approximate the densities $\pi(b_i \mid D)$ and $\pi(d_i \mid B)$ just once to make the generation simple and fast. For example, we could approximate the densities of the individual elements of b_i by piecewise normals and the density of d_i by piecewise gamma distributions. The priors for B and D can be derived from the prior for Σ either analytically or by simulation.

For example, suppose that we believe that the covariance structure is close to IID. We can use a prior that is centered on an IID covariance matrix. The errors in the difference form are $\varepsilon_{ij}^{*}-\varepsilon_{iJ}^{*}$, so that if the ε_{ij}^{*} are IID then Σ has diagonal elements equal to 1 and off-diagonal elements equal to 0.5. For $J=4$, Σ, B and D are

$$\Sigma_{IID}=\begin{bmatrix}1 & 0.5 & 0.5\\ 0.5 & 1 & 0.5\\ 0.5 & 0.5 & 1\end{bmatrix},\ B_{IID}=\begin{bmatrix}1 & 0 & 0\\ -1/2 & 1 & 0\\ -1/3 & -1/3 & 1\end{bmatrix} \text{ and } D_{IID}=\begin{bmatrix}1 & 0 & 0\\ 0 & 4/3 & 0\\ 0 & 0 & 1.5\end{bmatrix}$$

By centering the priors for B and D on B_{IID} and D_{IID} and using small variances, we can estimate a model close to IID probit. Making the priors more diffuse relaxes the IID assumption. If model identification is fragile as noted by Keane (1992), our approach allows estimates intermediate between IID and MNP models.

3.4 Empirical example

We now apply the sampling scheme in Section 3 to travel data collected by David Hensher and distributed with the Limdep Econometric Software. The same data were used by Chib, Greenberg and Chen (1998). The data come from a revealed preference experiment involving 210 subjects who chose among plane, train, bus and car modes for trips between cities in Australia. To allow comparisons with Chib et al., we selected four variables available to explain these choices: terminal waiting time (TWT), generalized cost (Cost), household income (Income) and size of travel party (PartySize). The latter two were assumed to only impact individual differences in the likelihood of choosing plane. Moreover three alternative specific constants (plane, bus and train; car = reference) were included. The model can be expressed as follows:

$$U_{\text{plane}} = \alpha_1 + \beta_1 TWT + \beta_2 Cost + \beta_3 Income + \beta_4 Partysize + \varepsilon_1$$

$$U_{\text{bus}} = \alpha_2 + \beta_1 TWT + \beta_2 Cost + \varepsilon_2$$

$$U_{\text{train}} = \alpha_3 + \beta_1 TWT + \beta_2 Cost + \varepsilon_3$$

$$U_{\text{car}} = \varepsilon_4$$

The following priors were used:

$$\beta \sim N(0, 10I_7),\ d_i \sim Gamma(d_{i,IID}; c_1, 1/c_1),\ b_i \sim N(b_{i,IID}, c_2 I_i),$$

where $c_1 = c_2 = 10$ and $d_{i,IID}$ and $b_{i,IID}$ are the values of d_i and b_i in the IID case as defined in Section 3. In further work not reported here we found that the prior for β can be made much less informative, e.g. $\beta \sim N(0, 10^6 I_7)$ without affecting the results, but the prior assumed above for the d_i and b_i cannot be. Estimates for IID probit were obtained by making c very small (e.g. $c_1 = c_2 = 10^{-10}$).

We studied the properties of the sampling scheme by running it for a warm-up period of 10,000 iterations, at the end of which we assumed that it was producing iterates from the correct posterior distribution, and a further 500,000 iterations to estimate the posterior means and standard deviations. Such a high number of iterations was feasible due to the small sample size and the speed with which our method generates draws from the covariance matrix. We wrote the program in Gauss, which required less than 2 hours, or less than 15 seconds per 1000 iterations (using one 625MHz processor on an Alpha Server 8400). The posterior means of β and Σ are presented in Tables 3.1 and 3.2 and are similar to those obtained by Chib, et. al. (1998).

Table 3.1: Posterior mean and standard deviation for β :

Variable	Mean	Standard deviation
Constant for plane	2.003	0.562
Constant for train	1.428	0.256
Constant for bus	1.189	0.247
Terminal waiting time	-0.030	0.007
Generalized cost measure	-0.010	0.002
Income * constant for plane	0.013	0.005
Party size * constant for plane	-0.459	0.112

Table 3.2: Posterior mean and standard deviation for Σ:

Elements of Σ	Mean	Standard deviation
σ_{12}	0.309	0.219
σ_{22}	0.540	0.261
σ_{13}	0.110	0.193
σ_{23}	0.215	0.138
σ_{33}	0.267	0.137

We note that the priors that we used on the covariance matrix were more dispersed than the priors used by Chib, et. al. (1998). We found that it was not possible to get sensible results if c_1 and c_2 are extremely large, for example $c_1 = c_2 = 10^{10}$, a value which effectively gives a flat prior. When we examined the priors used by other researchers, we realized that they had frequently assumed relatively tight prior information on the covariance matrix. Such informative priors may be necessary because of the fragile parameter identification noted by Keane (1992). We used Monte Carlo simulation (not reported here) to ensure that the sampling scheme was driven by the data rather than the priors even with c_1 and c_2 equal to 10. Our experience is that even with well-identified parameters, priors cannot be assumed to be completely flat. With experimental data, flatter priors (even with c_1 and c_2 equal to 10^3) generate sensible estimates but with $c_1 = c_2 = 10^{10}$ they do not.

Figure 3.1 and 3.2: Posterior mean estimates each over 1000 draws for σ12, σ22, σ23

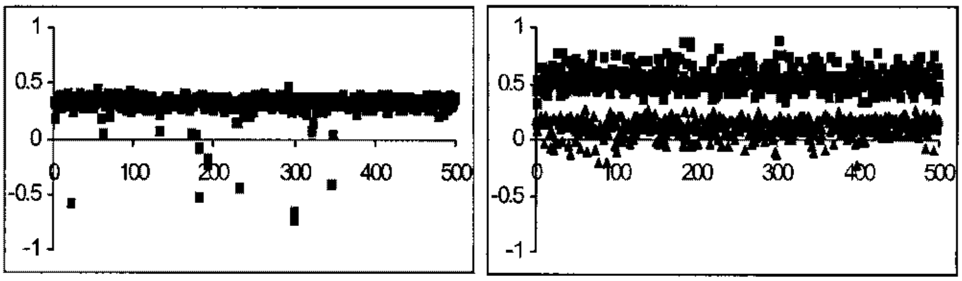
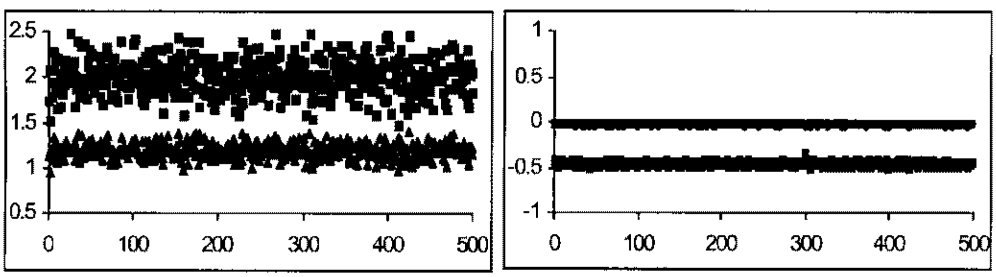


*Figure 3.3 and 3.4: Posterior mean estimates each based on 1000 draws for two alternative specific constants (Plane and Bus) and 2 other covariates (Terminal waiting time and Party size * constant for plane).*



Figures 3.1, 3.2, 3.3 and 3.4 graphically display the averages of the iterates (in blocks of 1000) of some of the parameters. These 500 averages can be taken as a sequence of posterior mean estimates each based on 1000 iterates.

It should be noted that we kept the range on the vertical axis constant to be able to compare the amount of variation in these plots. Clearly, there is much more variability in the distribution of the posterior mean estimates for the covariance elements and the alternative specific constants than the other covariates. The plot for σ_{12} in figure 3.1 is of most concern because it shows large outliers for some averages, suggesting that posterior mean estimates based on 1000 draws have a large left tail, which gets worse as c_1 and c_2 increase. Thus, informative priors are clearly needed, but how tight these priors need to be is an open question.

3.5 Selecting a subset of the covariance matrix elements

This section exploits the simplicity of the approach presented in section 3 to implement a selection procedure to reduce the number of elements necessary to parameterize the covariance matrix. This approach is motivated by the fragility reported in the estimation of the MNP model. Reducing the number of elements of the covariance matrix by selecting and estimating only the most important ones may lessen this problem. Moreover, the size of the covariance matrix increases with the number of choice options. With 4 choice options, it is necessary to estimate 5 elements of the covariance matrix, but with 10 options this number increases to 44. Therefore, to gain efficiency, one could try to estimate a sparser structure by selecting and estimating fewer elements.

To adapt the concept of variable selection in regression to selecting elements of the covariance matrix, or more precisely the decomposition of the inverse covariance, it is necessary to compare them to some base covariance matrix. An obvious base model is the IID structure, in which case the marginal posterior probability is computed that each element is equal to its IID counterpart. Instead of the base model being the identity, it is possible to use a more general diagonal structure. The procedure is unchanged but a diagonal covariance matrix is estimated in the first stage and used as the default for the selection procedure in the second stage.

To obtain identification in sections 2 to 4, the model was expressed in terms of differences from the last option. Identification can also be achieved by restricting the covariance matrix. For example, the variance and covariance elements of the last option can be restricted to 1 and 0's respectively. The selection can be done in either way, but this alternative approach seems simpler. Taking differences complicates the default model because any diagonal structure produces a full covariance matrix when the model is expressed as differences to a base; see section 3 for a representation of the IID structure in the difference space.

The model we now work with is written as follows:

$$U_{ij}^* = x_{ij}^{*}{}' \beta_j^* + \varepsilon_{ij}^*,$$

where ε_{ij}^* is $N\left(0,\Sigma^*\right)$, $\Sigma_{iJ}^*, \Sigma_{Ji}^* = 0$ for $i \neq J$ and $\Sigma_{11}^* = 1$.

The "*" notation was introduced in Section 2 to differentiate the model in its full representation from the differenced model used to achieve identification. To facilitate exposition, the "*" notation is now omitted as such differentiation is now unnecessary.

The covariance matrix is again decomposed into a lower triangular matrix and a diagonal one such that $\Sigma^{-1} = B'DB$. The flexibility of the method allows setting $\Sigma_{iJ}, \Sigma_{Ji} = 0$ and $\Sigma_{11} = 1$ by taking, without loss of generality, $B_{iJ}, B_{Ji} = 0$ and $D_{11} = 1$. We take a diagonal Σ as our default model which makes B the identity matrix in this model.

Let now $\gamma_i = (\gamma_{i,1}, \ldots, \gamma_{i,i-1})$ be a vector of indicator variables such that $\gamma_{i,j} = 0$ if b_{ij} is identically 0, and $\gamma_{i,j} = 1$ otherwise. In other words, at each iteration the indicator $\gamma_{i,j} = 1$ implies that element b_{ij} is estimated and $\gamma_{i,j} = 0$ indicate that b_{ij} must be fixed to zero.

The indicator variables $\gamma_{i,j}$ are assumed to be jointly independent a priori and have prior probabilities $\pi(\gamma_{i,j}=1)=0.5$ for all i and j. Using this notation the distribution of $\gamma_{i,j}$ can be obtained with D and B_γ's elements integrated out (as normal and gamma integrals, see section 3):

$$\pi\left(\gamma \mid \beta, D, B_\gamma, \delta\right) \propto \mid D\mid^{0.5N} \exp\left(-0.5 tr\, DB_\gamma AB_\gamma'\right)$$

The matrix A is defined and partitioned as in section 3, but since only the selected elements of B_γ need to be integrated out of the distribution, A_γ now consists of the corresponding selected elements in B_γ. The prior density for $b_{i\gamma_i} \mid D, \gamma, \delta$ is $N\left(b_{i\gamma}^0, c_{bi}\left(A_{i\gamma}d_i\right)^{-1}\right)$ and d_i is $gamma\left(d_i^0/c_{di}, 1/c_{di}\right)$.

By integrating $b_{i\gamma_i}$ out, we have:

$$\pi\left(\gamma_i \mid W, \beta\right) \propto d_i^{(N-q_{\gamma_i})/2} r^{q_{\gamma_i}/2} \mid A_{i\gamma_i}\mid^{-1/2} \exp\left(-1/2\left(\alpha_i - ra_{i\gamma_i}' A_{i\gamma_i}^{-1} a_{i\gamma_i} - a_{i\gamma_i}' b_{i\gamma_i}^0 r_1\right) d_i\right)$$

where $r = c_{di}/(1+c_{di})$, $r_1 = 1/(1+c_{di})$ and

q_{γ_i} represents the number of selected elements in $b_{i\gamma_i}$.

The d_i elements can be integrated out as a gamma integral:

$$\pi\left(\gamma_i \mid W, \beta\right) \propto \Gamma(\varphi) * r^{q_{\gamma_i}/2} \mid A_{i\gamma_i}\mid^{-1/2} \left(1/2\left(\alpha_i - ra_{i\gamma_i}' A_{i\gamma_i}^{-1} a_{i\gamma_i} - a_{i\gamma_i}' b_{i\gamma_i}^0 r_1\right) + 1/c_{di}\right)^{-\varphi}$$

where $\varphi = (N - q_{\gamma_i})/2 + d_i^0/c_{di}$

Since the vector γ_i is binary, conditional probabilities $\pi\left(\gamma_{ij}=1 \mid W, \beta, \gamma_{i,k\neq j}\right)$ and $\pi\left(\gamma_{ij}=0 \mid W, \beta, \gamma_{i,k\neq j}\right)$ must sum to 1 and can therefore be normalized and evaluated.

The Gibbs sampler can be used to generate γ_i. All other distributions can be sampled as in section 3. The method has been applied to the data described in section 4. Rather than reporting the mode of the estimates by estimating a final version of the model using a selected subset of variables, each draw of B_γ is used to compute Σ_γ which produces estimates that are averaged over all subsets/models considered.

Table 3.3 presents the diagonal structure used as default for the selection. It was obtained using an algorithm similar to that of section 3 but with an identity matrix in place of B and without taking differences with respect to a base alternative. 10,000 iterations were used for convergence and another 10,000 to generate both the default and the moments of the posterior distribution using the selection scheme.

Table 3.3: Posterior mean and standard deviation of the default covariance matrix

Default Σ	Mean	Standard deviation
σ_{11}	1	
σ_{22}	0.575	0.230
σ_{33}	0.199	0.168
σ_{44}	0.420	0.205

The results of the selection/averaging procedure generated with these default values are presented in tables 3.4 and 3.5. The priors are similar to the previous example expect for c which is made larger ($c_{di} = c_{bi} = 10; c = 10^6$). In conclusion, a diagonal matrix seems to be sufficient for these data. All covariance elements are clearly not significantly different from zero.

Table 3.4: Posterior mean and standard deviation for Σ

Elements of Σ	Mean	Standard deviation
σ_{12}	0.00136	0.042
σ_{22}	0.550	0.240
σ_{13}	-0.00118	0.029
σ_{23}	0.00405	0.031
σ_{33}	0.206	0.165

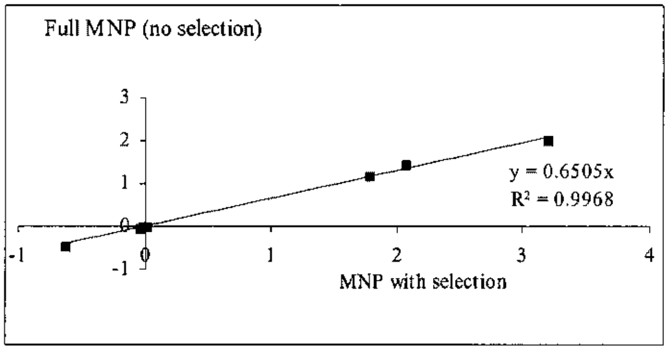
Note: σ_{44} is set to its default σ_{14}, σ_{24}, σ_{34} are zeroed for identification purposes

Table 3.5 presents the posterior means of the coefficients. To compare them to those presented in table 3.1, the fact that both models are not scaled identically must be taken into account. Effectively, since the first model is expressed as differences from a base while the second is not, setting $\Sigma_{11} = 1$ identifies the model in both cases but locates it differently. Therefore, there is a factor of proportionality between the two sets of estimates. Figure 3.5 demonstrates that relationship.

Table 3.5: Posterior mean and standard deviation for β *:*

Variable	Mean	Standard deviation
Constant for plane	3.202	0.556
Constant for train	2.073	0.245
Constant for bus	1.777	0.245
Terminal waiting time	-0.045	0.006
Generalized cost measure	-0.014	0.002
Income * constant for plane	0.015	0.007
Party size * constant for plane	-0.624	0.138

Figure 3.5: Posterior means estimates: Full MNP versus MNP with selection



The selection procedure also allows us to compute the probabilities that elements of the B matrix are different from zero. Table 3.6 presents these probabilities which confirm that a diagonal matrix is sufficient for these data.

Table 3.6: Probabilities that B's elements are different from zeros

Elements	Probabilities
B_{12}	0.121
B_{13}	0.080
B_{23}	0.123

Finally, we note that the selection scheme does not require much more processing, but this may not be obtained in problems with many more choice options. Selecting elements of the covariance matrix has implications for the priors. That is, their choice appears less critical, no doubt due to the fact that fewer elements had to be estimated in the example because a diagonal structure was sufficient. It also demonstrates the simplicity and flexibility of the method proposed in section 3.

3.6 Conclusions

This chapter introduced a new approach for Bayesian estimation of the MNP model. It allows for flexible priors for the covariance matrix, but is simple to apply. In view of the fragility of practical identification in MNP models, it is necessary to place informative priors on Σ, which makes the approach attractive. The flexibility of the method is further exploited by implementing a selection procedure on the elements of the covariance matrix, whose size may therefore be reduced if such a reduction is supported by the data. We considered a small example to show that the method is fast and produces results similar to previous MNP models estimated from these data. Moreover, the selection procedure demonstrates that a diagonal structure offers a good representation and a more complex error structure is probably not required for this specific example.

Chapter 4

Comparing MNL and MNP with Experimental Data: Are MNL Estimates Biased When Errors Are Non-IID?

4.1 Introduction

Choice models have been widely used in marketing to model consumer preferences. A reason for their popularity is the recognition that many marketing decisions involve choosing amongst mutually exclusive alternatives in real markets, including the choice to 'not choose'. Two widespread model forms are the multinomial logit (MNL) and multinomial probit (MNP); the latter allows for more general error structure. Staelin and Chapman (1982) pioneered the application of MNL models to scanner panel data; Louviere and Hensher (1983) and Louviere and Woodworth (1983) were the first to use them on choice experiment data. While it took longer for the more complex MNP to be applied, it now appears in the literature with increasing frequency (e.g., Chintagunta 1992; Bolduc, Lacroix and Muller 1996; Haaijer, Wedel, Vriens and Wandsbeek 1998).

Even though scanner panel data are now readily accessible, they are not the only choice data of interest to marketers. Another way of collecting information is provided by conjoint choice experiments, that now are widely used in marketing research to assess and predict the likely outcomes of new product introductions, changes to product configurations and many other applications (see Wittink and Cattin 1989). Louviere, Hensher and Swait (2000) discuss some advantages of choice experiments compared with real market choice data. For example, real market data do not allow one to estimate demand for new products with new features; they often lack variability in explanatory variables and are often highly collinear and expensive to collect. On the other hand, real market data do not suffer from interpretation biases that experiments might introduce. Therefore, many studies mix revealed and stated preference data as a joint enrichment strategy (e.g., Adamowicz, Louviere and Williams 1992; Ben-Akiva and Morikawa 1990; Swait, Louviere and Williams 1994; Hensher, Louviere and Swait 1999).

Real market and experimental data clearly differ but both MNL and MNP models are now used irrespective of the data source. Although models more complex than MNL such as Nested logit, heteroscedastic logit or mixed logit are now commonly used, much research has been based on MNL because historically it was the available tool with which to analyse discrete choice data. Not surprisingly, therefore, scholars have studied biases associated with the use of MNL when the true choice process is generated by another model such as MNP. For example, Horowitz (1980) assessed possible biases from using MNL when errors are not independently and identically distributed (IID); using revealed preference (RP) travel demand data he found that estimates from MNL models were consistent. However, questions remain about the performance of MNL estimated with stated preference (SP) data, and especially for different types of experiments in which errors are not IID.

We investigate MNP not out of a desire for more complexity but rather due to findings from previous studies in marketing. The IID error assumption does not in itself imply the independence of irrelevant alternatives (IIA) property of MNL, which states that a new alternative will impact the existing ones in proportion to their current market share. However, the core of the problem of interest in this chapter is in the assumption that the errors are independent. For example, as Ben-Akiva and Lerman (1985, p.109) note: "*...any model based on the assumption that all the disturbances are independent would necessarily yield counterintuitive forecasts for the red bus/blue bus problem*". As mentioned, many researchers have reported violations of this assumption, such as Simonson and Tversky (1992), who found that the same alternative can appear more or less attractive depending on other alternatives present in the choice set. This suggests that non-IID errors should be allowed when some brands in choice sets are more closely related than others because similar brands are likely to share unobserved attributes that affect their error terms in similar ways. Other examples include violations of the IIA property between "none/no purchase" options and other alternatives in the set (e.g., Bucklin and Lattin 1991; Chintagunta 1993; Dellaert, Borgers and Timmermans 1996). Such violations of the IIA property require models that are more complex than MNL to detect them. Since MNP models have the most general error structure, they allow researchers to detect any violation of assumptions associated with MNL models, even though simpler structures such as mixed logit models can also be useful in this regard.

Given that MNL assumes that each individual acts as if she satisfies the IIA axiom, and that there are reasons to believe that, in marketing, this assumption is not always appropriate, it is imperative to verify whether studies, which used MNL with choice experiments, made correct inferences when the errors were non-IID, or if all results should be reconsidered in the light of more general estimation procedures. Now that MNP can be estimated due to recent advances in optimisation methods and more powerful computers, the gains from shifting from MNL to MNP may be examined more easily. It is also of interest to make sure that experimental data provide enough information to estimate the structure of the covariance matrix, which is the real benefit of MNP.

This chapter examines MNL and MNP estimation for choice experiments with non-IID errors. To do this, we use Monte Carlo (MC) simulation with two different types of experimental designs. The remainder of the chapter is organised as follows: in the next section, MNL and MNP models are briefly reviewed followed by a description of the designs investigated. Section 4 presents the results of the MC simulation. Conclusions and directions for further research complete the discussion.

4.2 Choice Models

MNL and MNP models are increasingly used to analyse choice (multinomial response) data when subjects choose one, and only one, option among a set of mutually exclusive alternatives. In this section, both models are introduced in greater detail.

MNL and MNP can be derived from Random Utility Theory (RUT). Consumers are assumed to associate a utility (U_{ij}^*) with each alternative in a choice set and are assumed to choose the most preferred alternative as if they are using a utility maximising behavioural rule. It is further assumed that the utilities are linear functions of the attributes (x_{ij}) that describe the alternatives. These model assumptions lead to the following expression:

$$U_{ij}^* = \alpha_j^* + x_{ij}'\beta_j^* + \varepsilon_{ij}^* \quad (4.1)$$

The indirect utilities U_{ij}^* for individual i and alternative j ($j = 1,2,...,J$) are unobserved variables since only discrete choices can be monitored. β_j^* is a vector of coefficientss that represent consumers' trade-offs among different attributes of a specific product (and captures what is known as the systematic component of utility). It is important to note that the coefficients can be alternative-specific (as indicated by a subscript for β_j^*), or generic (as indicated by no subscript for β^*). α_j^* represents a vector of alternative-specific constants (ASC's). Finally, a random component ε_{ij}^* is introduced to reflect that the analyst does not understand the choice process perfectly.

Although the U_{ij}^{*}'s are unobserved, given that consumers are assumed to choose the alternative they prefer most, the following is known:

$U_{ij} = 1$ if $U_{ij}^{*} > U_{ik}^{*}$ for any $j \neq k$,

$U_{ij} = 0$ otherwise.

In other words, U_{ij} is an indicator variable, which takes the value 1 if alternative j has been selected and 0 otherwise.

Models in which the dependent variable only satisfies ordinal properties have two identification issues that can be appreciated by recognizing that if one adds or multiplies equation (4.1) by a positive constant the indicator variable is unaffected. In other words, utilities can be rescaled by any order-preserving transformation, which leaves the indicator unaffected. These two identification issues are typically dealt with by setting the utility of one alternative to zero (i.e. the J's one) and restricting one element of the covariance matrix (i.e. element 1,1). This can be accomplished, for example, by first subtracting U_{iJ}^{*} from U_{ij}^{*} as follows,

$$W_{ij} = U_{ij}^{*} - U_{iJ}^{*} = \alpha_j^{*} + x_{ij}^{*}{}'\beta_j^{*} - \alpha_J^{*} - x_{iJ}^{*}{}'\beta_J^{*} + \varepsilon_{ij}^{*} - \varepsilon_{iJ}^{*} = \alpha_j + x_{ij}'\beta_j + \varepsilon_{ij}$$

which in vector notation becomes: $W_i = X_i\beta + \varepsilon_i$ where $\varepsilon_i \sim N(0, \Sigma)$, and by secondly setting $\Sigma_{11} = 1$. The result of these identification restrictions is that only differences between utilities can be estimated.

Different assumptions about the error distribution allow one to derive the MNL or MNP models. For example, McFadden (1973) showed that if the error distribution is IID

Gumbel (or equivalently type I extreme value), the parameters of interest can be estimated with a MNL model. McFadden accomplished this by extending the original paired comparison work of Thurstone (1927) to the multiple choice case. The Multinomial Probit (MNP) model introduced earlier is obtained if one assumes a normal distribution for the random components. The real advantage of MNL over MNP, and the reason it has been more commonly used, is primarily computational because a closed form expression for the probabilities exists, which allows one to quickly locate the Maximum Likelihood estimates.

Recently, advances have been made in estimation of MNP. For example, to use Maximum Likelihood (ML) to estimate MNP requires the evaluation of multiple integrals to the compute probabilities. Obviously, as the number of choice alternatives increases, the computational burden associated with multiple integrals also increases. A solution for this problem is to simulate probabilities and therefore evaluate the log likelihood (LL) for a particular set of parameters. One approach is the Geweke-Hajivassiliou-Keane simulator (GHK), which gives very quick and accurate approximations of the probabilities (i.e. Geweke 1991, Hajivassiliou 1994 and Keane 1994). Although the simulated probabilities produced by the GHK are unbiased, they do not yield an unbiased LL (Börsch-Supan and Hajivassiliou 1993). However, in practice Simulated Maximum Likelihood (SML) has been shown to work well (Geweke, Keane and Runkle 1994). Recently, the precision of SML method has been improved by the introduction of Halton sequences (Train 1999; Revelt and Train 1999).

Another approach to estimating MNP models is based on a Bayesian framework (e.g., McCulloch and Rossi 1994; Chib, Greenberg and Chen 1998), that uses the data augmentation technique detailed in Albert and Chib (1993). Chapter 3 discusses this in more details.

When comparing different estimation procedures, it is important to bear in mind that the variance of the Gumbel distribution is equal to $\pi^2/6\sigma^2$ (Ben Akiva and Lerman 1985). Therefore, MNP and MNL estimates should be closely related but require rescaling to achieve this. In a case of IID errors, MNL and Identity probit (i.e., the Identity probit model is defined as the MNP model assuming an IID error structure) should be equal up to a rescaling factor that is approximately equal to 1.6 ($=\pi^2/6$ - see Amemiya 1981). This chapter considers non-IID errors; hence even if all the models considered produced unbiased estimates, parameters would not be equal but proportional. Unfortunately the constant of proportionality that rescales MNL and Identity probit parameter estimates to non-IID MNP is unknown. Therefore, to compare the three models, all estimates are re-centered (or rescaled) around their true values. We assume that MNP is the true model; thus it should not need to be rescaled; however, since MNP cannot be centered perfectly due to sampling errors, centering the other two models could give them an unfair advantage. Despite the latter observation, we find that the rescaling factor in the case of MNP is very close to 1 (around 0.96). To assess differences in standard deviations the same rescaling factor must be applied, but t-ratios can be more easily compared because they are not affected by the scaling constant.

4.3 Conjoint Experiments

Typically, choice experiments consist of a series of designed scenarios (or choice sets), and each choice set describes several competing products. A sample of respondents chooses their preferred product from each choice set. Scenarios are usually constructed from fractional factorial designs whose objective is to estimate the impact of each attribute on the respondent's decision as independently as possible of all other attributes.

Many studies (e.g., Anderson and Wiley 1992; Kuhfeld, Tobias and Garrat 1994; Bunch, Louviere and Anderson 1996; Huber and Zwerina 1996) have investigated the problem of how to devise the most statistically efficient experimental designs for choice models. Although orthogonality does not ensure efficiency, it is usually desirable, but other criteria that have been investigated for constructing these kind of experiments include level balance and minimal overlap.

It is worth noting that the reasons for asking only choices from respondents have been well documented. That is one may wonder why experimentalists would construct tasks that yield only discrete information (such as choices) instead of observing ratings or rankings which appear to contain more information and/or are associated with easier estimation procedures. In particular, Louviere, Hensher and Swait (2000) explain that ratings and rankings require strong assumptions about human cognitive abilities to be satisfied in order for response to be of a higher measurement level than discrete choices. Moreover, choice tasks are more realistic than rating or ranking tasks and the issue of differences in response scales and scale usage between individuals does not arise (e.g., Elrod, Louviere and Davey 1992; Carroll and Green 1995). In addition, discrete choice tasks are quite easy and can be completed fairly quickly by respondents; hence respondents can evaluate more profiles. For example, Johnson and Orme (1996) showed that at least 20 choice sets could be administered to subjects without much decrease in information quality. Finally, choice experiments easily produce market share estimates, which are often useful to practitioners, without the need for ad hoc assumptions to translate stated intentions or conjoint model predictions into choices (Louviere 1988).

There are a very large number of possible methods to design choice experiments that could benefit from an investigation of the impact of using MNL when the true underlying process involves non-IID errors. Thus, we restrict our investigation to commonly used designs involving five attributes and four alternatives (one of which is a constant) and consider main effects only designs. The first experimental design was constructed following the approach described in Bunch et al. (1996), who showed that shifted designs (also called "cyclic designs" by Huber and Zwerina 1996) were the most efficient for the MNL estimation of main effects only for the case of generic coefficient specifications. Shifted designs use an orthogonal array to construct the first alternative and shift its levels using modular arithmetic to construct additional alternatives. In order to be able to construct three alternatives from an original orthogonal array, at least three levels for each attribute are needed; hence we used a 4^{5-3} orthogonal main effects design (in 16 profiles) to construct the first alternative, and shifted it twice to obtain two additional alternatives.

Because properties of experimental designs for choice tasks are not yet well known, there is little guidance available to construct designs that are consistent with non-IID models. On the other hand, experimental designs that result in attribute columns that are orthogonal within- and between-alternatives have been used to estimate non-IIA or cross-effects models. For example, Kuhfeld, Tobias and Garratt (1994) suggested that it is standard practice for choice experiments to be designed such that all main effects are estimable within- and between-alternatives. Such designs are also useful for estimating alternative-specific coefficient models because they ensure that the attribute columns belonging to different alternatives are also orthogonal to each other. The preceding suggests that it would be of academic and commercial interest to test two designs: (1) as discussed above, a design that is orthogonal only within-alternatives (e.g., a 4^{5-3} fractional factorial), which is consistent with a generic coefficient MNL model (but not alternative-specific coefficient models); and (2) a design that is orthogonal within- and between-alternatives (15 attribute columns taken from a 4^{21-18} fractional factorial). To estimate main effects only of 15 attributes in four levels, a design in 64 profiles is selected; that is the 4^{21-18}. The literature suggests that only the latter design, should allow one to precisely estimate non-generic coefficients; thus a key research question addressed in this chapter is whether either design can retrieve the correct parameters if the errors are non-IID. A second research question is whether MNL models yield biased estimates for both designs if errors are not IID. Both research questions are studied under the assumptions of generic and alternative-specific coefficients. The conditions investigated are summarised in Table 4.1 and the research questions can be expressed as differences between conditions A to H. More precisely, the first research question investigates the following comparisons: A vs. C, B vs. D, E vs. G and F vs. H. The second research question compares A vs. E, B vs. F, C vs. G and D vs. H.

Table 4.1: Conditions investigated under non-IID errors assumption

	MNL		MNP	
	Generic	**Alt.-spec.**	**Generic**	**Alt.-spec.**
4^{5-3}	A	B	E	F
4^{21-18}	C	D	G	H

A constant alternative was added for all choice sets since a constant alternative like "none" or a no-purchase option is often included in practice to make experiments closer to reality (Carson et al. 1994), among other reasons. Such options allow respondents to indicate that if the described alternatives were their only options they would not choose any. Another form of constant alternative is a "your current" or "latest chosen brand" option, which allows respondent to switch from their current product to one offered by the experiment. The choice of a constant alternative depends on the application and example of each can be found in Elrod et al. (1992) for an "own" constant alternative or in Oppewal, Louviere and Timmermans (1994) for a "none" option. The type of base alternative is not relevant to the research in this paper thus all attributes describing this alternative were coded zero, which is equivalent to a "none" option. Using base options can also be important from a statistical efficiency perspective. That is, as explained previously, choice models only estimate differences between alternatives; hence adding a constant alternative ensures that differences between the base and all other alternatives remain orthogonal (Dellaert, Borgers and Timmermans 1996 also point to this). This is not always important because clearly an experiment which is orthogonal between and within-alternatives would not require the constant option to ensure that the differences remain orthogonal. It can nevertheless be important depending on the strategy chosen to construct the experiment. Not withstanding the orthogonality issue, optimal designs cannot be obtained by only ensuring that all differences are orthogonal because the information matrix used to define optimality criteria such as D-optimality or A-optimality is a function of the unknown model estimates. Thus, optimal designs are difficult to obtain and can apparently be attained only after the estimates are known. Not surprisingly, therefore, most researchers (e.g., Bunch et al. 1996; Street et al. 1999) have investigated design optimality assuming the null hypothesis of no difference between the effects of each attribute to be true (i.e. model estimates equal zero).

It is also important to check how the estimated covariance matrix compares with the true values for each choice experiment. For example, Dellaert, Borgers, Louviere and Timmermans (1996) argue that designs in which attributes are orthogonal only within-alternatives, as well as designs in which attributes are orthogonal within- and between-alternatives, may be inefficient if there are IID violations or may have identification problems. They propose a method that uses interrelated designs to identify and estimate models with non-IID errors. Thus, the literature is unclear as to whether one can estimate the structure of the covariance matrix properly with the experiments considered in this chapter.

The particular issues addressed in this chapter now can be more clearly stated. That is, we study the extent of biases that arise if one estimates choice experiment effects with

MNL models but the true errors are not IID. The following experimental situations are investigated:

1. Coefficients are assumed to be generic and the design is orthogonal only within-alternatives (4^{5-3}).
2. Coefficients are assumed to be alternative-specific and the design is orthogonal only within-alternatives (4^{5-3}).
3. Coefficients are assumed to be generic and the design is orthogonal within- and between-alternatives (15 variables from 4^{21-18}).
4. Coefficients are assumed to be alternative-specific and the design is orthogonal within- and between-alternatives (15 variables from 4^{21-18}).

Identity probit models are also estimated; we expect Identity probit to produce parameters very similar to MNL due to similar assumptions about error terms. The distribution of the true model is normal; hence any biases from using MNL can be attributed to the different underlying distribution rather than to a non-IID error assumption. Including the Identity probit in the research allows us to determine if it manifests the same biases as MNL; and if there are biases, it will be clear that they are due to assumptions about the covariance matrix rather than the distribution per se (i.e. Gumbel vs. Normal). Moreover, it will rule out any potential differences due to different methods of estimation. That is, MNP and Identity probit are estimated using bayesian methods but maximum likelihood is used to estimate MNL. The MNP model is estimated using the approach discussed in Chapter 3. As Revelt and Train (1999) argue, the choice of the method is mostly a case of what the researcher feels more comfortable with.

4.4 Results

This section presents results from Monte-Carlo simulations (30 replications) using MNL, Identity probit and MNP models to estimate model parameters from choice experiments when errors are not IID. The covariance matrix of the error structure used in these simulations is presented in Table 4.2. It was chosen to be clearly different from the IID structure.

Table 4.2: Covariance matrix for IID errors vs. covariance matrix for non-IID errors

$$\begin{bmatrix} 1 & -0.5 & -1 \\ -0.5 & 2.5 & -2.5 \\ -1 & -2.5 & 6 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}$$

It should be noted that all the models in this research estimate attribute main effects only.

Case A vs. E: Generic coefficients with a 4^{5-3} design

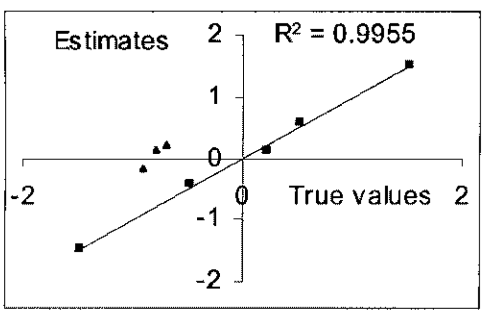
The first results in Table 4.3 pertain to applying the three models to the smallest experimental design (4^{5-3}) assuming that coefficients are generic. That is, the true model is generic but was generated with non-IID errors. The average parameter estimates, rescaled for the MNL and Identity probit to be comparable to MNP, are shown in Table 4.3. The parameters used for these simulations are displayed in the column "Truth" of Table 4.3.

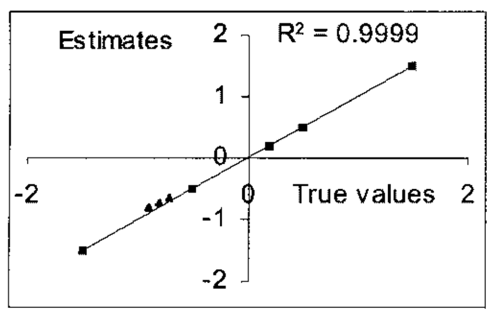
Table 4.3: Average parameter estimates and biases (generic coefficients) with 4^{5-3}

Labels	Truth	Average parameter estimates			Average biases		
		MNL	Identity probit	MNP	MNL	Identity probit	MNP
Constant 1	*-0.9*	*-0.193*	*-0.083*	*-0.813*	*0.707*	*0.817*	*0.087*
Constant 2	*-0.8*	*0.131*	*0.328*	*-0.715*	*0.931*	*1.128*	*0.085*
Constant 3	*-0.7*	*0.216*	*0.477*	*-0.658*	*0.916*	*1.177*	*0.042*
Coefficient 1	-0.5	-0.401	-0.442	-0.496	0.099	0.058	0.004
Coefficient 2	-1.5	-1.466	-1.468	-1.505	0.034	0.032	-0.005
Coefficient 3	0.2	0.149	0.142	0.191	-0.051	-0.058	-0.009
Coefficient 4	0.5	0.590	0.609	0.483	0.090	0.109	-0.017
Coefficient 5	1.5	1.529	1.510	1.503	0.029	0.010	0.003

The first column contains the true parameters, which are constants (α_i) or attribute coefficients (β's) in the design. Not surprisingly, the MNL and Identity probit results are similar, in terms of estimates of constants and coefficients, and thus their biases are also similar. However, more interesting is the fact that their coefficient estimates (β's) are very close to the true values, and in fact, are almost as close to the true values as the MNP estimates despite the fact that MNP is the correct model for these data. This finding is not a complete surprise since Horowitz (1980) previously noted that MNL produces consistent estimates when errors are not IID. It is worth noting, however, that the covariance matrix used in these simulations represents a severe departure from IID assumption. Importantly, there seems to be a consensus in the literature that MNL estimates will be biased if the IID error assumption is violated, and it is also believed that designs such as the ones used in this chapter will not efficiently estimate the true model parameters (Dellaert et al. 1996; Vriens et al. 1998). Table 4.3 shows, however, that the difference between the two non-IID models and MNP is in the constants (ASC's). The coefficient estimates (β's) produced by MNL are so close to those recovered by MNP that if the objective of the study were only about estimating and understanding the attribute trade-offs, there would be not much from estimating MNP even if errors are not IID. Figures 4.2 and 4.3 illustrate this by graphing the true values against the average parameter estimates of MNL and MNP respectively.

Figures 4.2 and 4.3: MNL and MNP avg. estimates against true values (with 4^{5-3})





▲ = *ASC's and* ■ = β's

The graphs in Figures 4.2 and 4.3 allow one to visually assess the bias in the constants. The fitted line used only the coefficients as data, and the R-squared value is a good measure of how well MNL recovers the true β values. It is worth noting that the R-squared value for the MNP estimates vs. true values is only marginally better, which supports the contention that MNL recovers the true parameters in this case.

In the case of predictive validity, however, the performance of MNL depends on how much the true error structure differs from IID. Table 4.4 shows the average log likelihood at solution obtained for each model (GHK simulator was used in the case of

MNP), which clearly indicates that MNP outperforms the other two models. Thus, MNP produces more accurate within sample predictions.

Table 4.4: Average log likelihood at solution (generic coefficients with 4^{5-3})

Average log likelihood at solution		
MNL	Identity probit	MNP
-1255.77	-1247.19	-1177.95

Comparing standard deviations is also important to fully assess the performance of MNL in this case. Table 4.5 contains the average standard deviations and t-ratios. As explained earlier, standard deviations were rescaled using the same factor as applied to the coefficients (β 's) to facilitate comparisons.

Standard deviations for both IID models are somewhat smaller than those for MNP, but a comparison of t-ratios may be more informative. Specifically, the large t-ratios (>10) associated with the IID models are much too large compared to those associated with MNP, although, the smaller t-values (<10) are similar for all three models. Practically speaking, however, this suggests that the right conclusions would be reached based on the IID models, as only already large t-ratios seem to be overestimated. That is, attributes that are very important would be inferred to be even more important based on the IID models. However, this is one simulation result, and more work is needed to determine its generality. Finally, this case shows that constants (ASC's) are biased for both IID models, so their t-ratios cannot be compared with those estimated from MNP.

Table 4.5: Average standard deviations and t-ratios (generic coefficients with 4^{5-3})

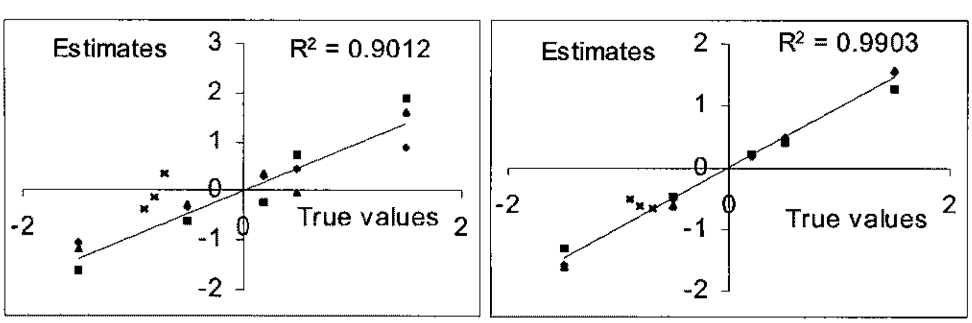
Labels	Average standard deviation			Average t-ratios		
	MNL	Identity probit	MNP	MNL	Identity probit	MNP
Constant 1	*0.207*	*0.204*	*0.145*	*0.93*	*0.41*	*5.60*
Constant 2	*0.232*	*0.225*	*0.198*	*0.57*	*1.46*	*3.61*
Constant 3	*0.247*	*0.237*	*0.195*	*0.88*	*2.01*	*3.37*
Coefficient 1	0.052	0.052	0.065	7.72	8.43	7.64
Coefficient 2	0.071	0.063	0.121	20.69	23.17	12.41
Coefficient 3	0.052	0.051	0.053	2.87	2.78	3.61
Coefficient 4	0.053	0.053	0.051	11.15	11.60	9.38
Coefficient 5	0.076	0.067	0.122	20.21	22.70	12.31

On the whole, the results of this first case are encouraging because research in consumer behaviour in marketing research applications often assumes generic coefficients. The results suggest that if coefficients are generic, conclusions about trade-offs in consumer preferences are invariant to assumptions placed on error terms. Thus, in this type of applications, (e.g. Yeoh et al. 1998) predictions are of less interest, which offsets the one major advantage of MNP. It is also interesting to note that in studies which compare different types of experimental conditions, MNL models are often used to estimate a common coefficient vector from all sets of data assuming that there are different error variances for different data sources (e.g., Brazell and Louviere 1998; Brydon et al. 1999; Diener and Dellaert 1997; Louviere et al. 1993; Severin et al. 1999; Yeoh et al. 1998;). Failure to reject the null hypothesis when pooling sources of preference data implies that differences in estimates can be attributed to differences in choice consistency, not differences in means. Typically, such experiments assume that constants can be allowed to differ but coefficients should be rescaled. The present results suggest in contrast that the need to allow for different constants for each data source may be a consequence of the fact that a more general error covariance matrix underlies each data set, and this error structure only impacts the constants. If it can be shown to generalise, this finding also would explain why such studies demonstrate very often stability in means, because if all MNL estimates were biased due to non-IID errors, one would presume that stability would be rejected much more often.

Case B vs. F: Alternative-specific coefficients with a 4^{5-3} design

We now consider a second and a priori less interesting case that examines the implications of estimating the same models as before but for the case of alternative-specific coefficients (β's) based on data devised from an experiment which a priori should not be efficient for investigating this more general structure (4^{5-3}). We readily admit that such a model analysis is not likely to be undertaken on purpose, but we can imagine that researchers may be tempted to extract more information from the data in situations in which there are limited resources available or initial results are unexpected. Thus, we believe that it is of interest to evaluate how MNL or MNP would perform in such a case. The results are in Table 4.3b, and summarised graphically in Figures 4.2b and 4.3b for MNL and MNP.

Figure 4.2b and 4.3b: MNL and MNP avg. estimates against true values (with 4^{5-3})



▲ = β_1's *(alt.1)*, ■ = β_2's *(alt.2)*, ♦ = β_3's *(alt.3)*, × = ASC's

Table 4.3b: Average parameter estimates and biases estimating assuming alternative-specific coefficients with a 4^{5-3}.

Labels	Truth	Average parameter estimates			Average biases		
		MNL	Identity probit	MNP	MNL	Identity probit	MNP
Constant 1	*-0.9*	*-0.353*	*-0.250*	*-0.502*	*0.547*	*0.650*	*0.398*
Constant 2	*-0.8*	*-0.149*	*0.073*	*-0.631*	*0.651*	*0.873*	*0.169*
Constant 3	*-0.7*	*0.371*	*0.718*	*-0.653*	*1.071*	*1.418*	*0.047*
Coefficient 1 of alt1	-0.5	-0.604	-0.595	-0.461	-0.104	-0.095	0.039
Coefficient 1 of alt2	-0.5	-0.286	-0.369	-0.624	0.214	0.131	-0.124
Coefficient 1 of alt3	-0.5	-0.325	-0.332	-0.503	0.175	0.168	-0.003
Coefficient 2 of alt1	-1.5	-1.617	-1.642	-1.308	-0.117	-0.142	0.192
Coefficient 2 of alt2	-1.5	-1.206	-1.304	-1.624	0.294	0.196	-0.124
Coefficient 2 of alt3	-1.5	-1.049	-1.063	-1.577	0.451	0.437	-0.077
Coefficient 3 of alt1	0.2	-0.199	-0.154	0.216	-0.399	-0.354	0.016
Coefficient 3 of alt2	0.2	0.366	0.343	0.191	0.166	0.143	-0.009
Coefficient 3 of alt3	0.2	0.310	0.310	0.178	0.110	0.110	-0.022
Coefficient 4 of alt1	0.5	0.760	0.776	0.381	0.260	0.276	-0.119
Coefficient 4 of alt2	0.5	-0.036	0.091	0.478	-0.536	-0.409	-0.022
Coefficient 4 of alt3	0.5	0.470	0.416	0.485	-0.030	-0.084	-0.015
Coefficient 5 of alt1	1.5	1.879	1.843	1.285	0.379	0.343	-0.215
Coefficient 5 of alt2	1.5	1.595	1.604	1.582	0.095	0.104	0.082
Coefficient 5 of alt3	1.5	0.876	0.889	1.552	-0.624	-0.611	0.052

In this case, all models experience difficulty in estimating the true coefficients and constants, including MNP, which places the correct assumption on the covariance matrix. However, it is worth noting that MNP does a good job given the small experiment (4^{5-3}) that was used. Indeed, it appears that this case exposes some important issues from a design or data point of view. That is, claims that non-IID errors produce biased estimates appear to be correct for this case, but again, this design would not have been selected on purpose to estimate such an error structure, so the results are still surprisingly good. It is likely that these biases arise because the attributes are not orthogonal within- and between-alternatives in this case. It also seems clear that MNL and Identity probit cannot recover the alternative-specific coefficients, and although

MNP is better, it is not perfect. On the other hand, the results also suggest that alternative-specific coefficients (β's) are affected in different proportions, but it may be due to the design which is not efficient for alternative-specific estimates. Figures 4.2b and 4.3b show that the coefficients belonging to an alternative seem to be affected similarly (this is more for MNL), but the constants (represented by small crosses), appear to be more biased for MNL than MNP.

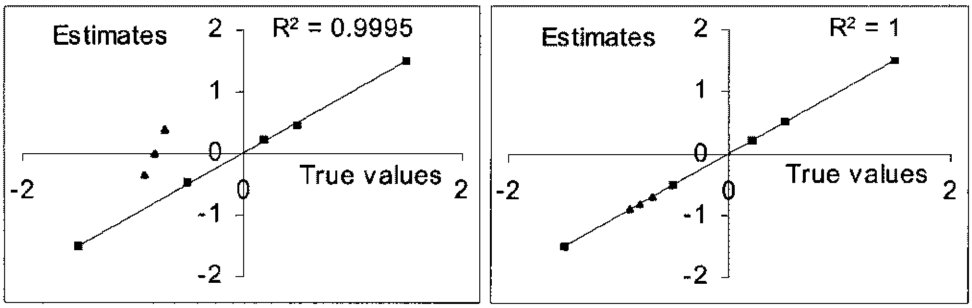
In conclusion, if alternative-specific coefficients are of interest, designs which are only orthogonal within-alternatives should be used in conjunction with MNP because MNP appears to produce more accurate estimates than MNL (or Identity probit).

However, although MNP estimates are better in this case, they do not appear to be unbiased. An obvious question for future research is whether a larger initial fractional design would reduce the bias in MNP significantly, or whether the bias observed is larger for the shifted designs regardless of design size.

Case C vs. G: Alternative-specific coefficients with a 4^{21-18} design

We now investigate whether a better design, namely one which is orthogonal within- and between-alternatives will perform better under the same conditions. Based on our earlier results, it would seem that if a researcher only seeks generic coefficients, between-alternatives orthogonality might be unnecessary. Nonetheless, it is worth investigating whether ASC's associated with MNL might be unbiased or less biased if attributes are orthogonal within- and between-alternatives. We briefly discuss this result by considering Figures 4.2c and 4.3c.

Figure 4.2c and 4.3c: MNL and MNP avg. estimates against true values (with 4^{21-18})



▲ = *ASC's and* ■ = β's

That is, the same conclusions apply as for the smaller design, namely that MNL estimates the coefficients accurately but the constants (ASC's) are biased. Table 4.3c can be compared to Table 4.3 to see how within- and between-orthogonality improves the model estimates.

Table 4.3c: Average parameter estimates and biases (generic coefficients with 4^{21-18})

Labels	Truth	Average parameter estimate			Average bias		
		MNL	Identity probit	MNP	MNL	Identity probit	MNP
Constant 1	*-0.9*	*-0.351*	*-0.283*	*-0.901*	*0.549*	*0.617*	*-0.001*
Constant 2	*-0.8*	*0.010*	*0.101*	*-0.806*	*0.810*	*0.901*	*-0.006*
Constant 3	*-0.7*	*0.390*	*0.516*	*-0.689*	*1.090*	*1.216*	*0.011*
Coefficient 1	-0.5	-0.473	-0.465	-0.511	0.027	0.035	-0.011
Coefficient 2	-1.5	-1.498	-1.495	-1.497	0.002	0.005	0.003
Coefficient 3	0.2	0.229	0.238	0.208	0.029	0.038	0.008
Coefficient 4	0.5	0.473	0.471	0.503	-0.027	-0.029	0.003
Coefficient 5	1.5	1.514	1.519	1.498	0.014	0.019	-0.002

In the case of generic coefficients, between orthogonality does not seem to improve the efficiency of the estimates by much, especially considering that 64 profiles are used for this experiment. Thus, is there a significant gain in moving to designs that require many more profiles (e.g., 4 times as many) when estimating generic coefficients? An important issue, not yet discussed, concerns the estimated covariance matrix; that is, MNP imposes a more general error structure, which must be estimated. Therefore, even if there are some gains in using MNP over MNL, it is important to investigate how well can MNP estimate the error covariance matrix for each experimental design.

As discussed previously, there is legitimate concern about whether the designs considered in this research can estimate the elements of the covariance matrix efficiently, especially for the small number of respondents assumed in these simulations (100). Table 4.6 presents these estimates for both experiments assuming generic coefficients.

Table 4.6: Average estimates of the covariance matrix assuming generic coefficients

Labels	True values	Average estimates of the covariance matrix	
		Using a 4^{5-3} design	Using a 4^{21-18} design
Sigma 11	1	1	1
Sigma 21	-0.5	-0.428	-0.582
Sigma 22	2.5	2.962	2.901
Sigma 31	-1	-1.176	-1.109
Sigma 32	-2.5	-2.912	-2.831
Sigma 33	6	7.545	7.030

Although the estimates do not equal the true values, they are surprisingly close to the true covariance elements. The results also suggest that between orthogonality may increase the efficiency of the estimates of the covariance matrix; however, because there are four times as many profiles in this design, the apparent increase in efficiency may be due to sampling from a broader space in the design associated with the larger number of combinations. Future research should address this more general question, but our results suggest that for generic coefficients, *MNP seems to provide a good representation of the covariance structure even with small numbers of respondents and main effects only designs in which attributes are only orthogonal within-alternatives.*

This is a very encouraging result because studies which can assume generic coefficients may use MNP with less profiles than the popular between and within-orthogonality design.

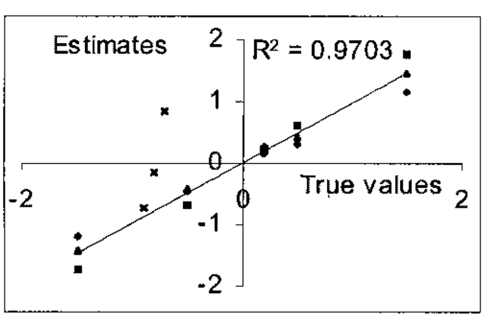
Case D vs. H: Alternative-specific coefficients with a 4^{21-18} design

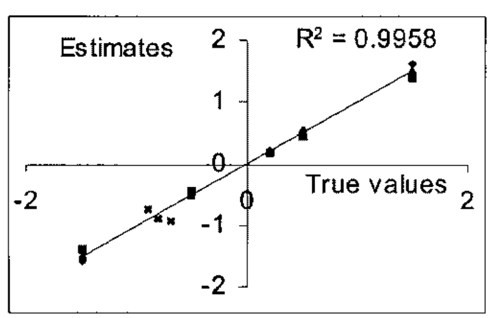
The last set of results concern the estimation of alternative-specific coefficients from designs, which should allow them to be estimated because all attributes are orthogonal within- and between-alternatives. However, as before, the design that we used does not account for non-IID errors, and so one should again expect to observe some bias. It is therefore interesting to know whether the same conclusions as for generic coefficients apply. Table 4.3d contains the average parameter estimates for the three models considered, estimated from non-IID simulated data; Figures 4.1d and 4.2d present the same results graphically.

Table 4.3d: Average parameter estimates and biases estimating assuming alternative-specific coefficients with 15 variables from a 4^{21-18}.

Labels	Truth	Average parameter estimates			Average biases		
		MNL	Identity probit	MNP	MNL	Identity probit	MNP
Constant 1	*-0.9*	*-0.713*	*-0.623*	*-0.759*	*0.187*	*0.277*	*0.141*
Constant 2	*-0.8*	*-0.142*	*-0.038*	*-0.914*	*0.658*	*0.762*	*-0.114*
Constant 3	*-0.7*	*0.841*	*0.859*	*-0.929*	*1.541*	*1.559*	*-0.229*
Coefficient 1 of alt1	-0.5	-0.709	-0.696	-0.446	-0.209	-0.196	0.054
Coefficient 1 of alt2	-0.5	-0.434	-0.438	-0.515	0.066	0.062	-0.015
Coefficient 1 of alt3	-0.5	-0.444	-0.421	-0.522	0.056	0.079	-0.022
Coefficient 2 of alt1	-1.5	-1.712	-1.720	-1.372	-0.212	-0.220	0.128
Coefficient 2 of alt2	-1.5	-1.412	-1.414	-1.508	0.088	0.086	-0.008
Coefficient 2 of alt3	-1.5	-1.204	-1.179	-1.585	0.296	0.321	-0.085
Coefficient 3 of alt1	0.2	0.240	0.242	0.182	0.040	0.042	-0.018
Coefficient 3 of alt2	0.2	0.257	0.262	0.223	0.057	0.062	0.023
Coefficient 3 of alt3	0.2	0.141	0.147	0.225	-0.059	-0.053	0.025
Coefficient 4 of alt1	0.5	0.613	0.605	0.450	0.113	0.105	-0.050
Coefficient 4 of alt2	0.5	0.437	0.444	0.536	-0.063	-0.056	0.036
Coefficient 4 of alt3	0.5	0.320	0.329	0.528	-0.180	-0.171	0.028
Coefficient 5 of alt1	1.5	1.781	1.793	1.374	0.281	0.293	-0.126
Coefficient 5 of alt2	1.5	1.445	1.439	1.500	-0.055	-0.061	0.000
Coefficient 5 of alt3	1.5	1.157	1.161	1.616	-0.343	-0.339	0.116

Figures 4.1d and 4.2d: MNL and MNP avg. estimates against true values (with 4^{21-18})





▲ = β_1's *(alt.1)*, ■ = β_2's *(alt.2)*, ♦ = β_3's *(alt.3)*, × = ASC's

The results suggest that failure to take non-IID errors into account in the estimations seems to definitely impact the constants (ASC's); that is, MNL and Identity probit again produce biased estimates of the ASC's. In the case of coefficients (β's), this design improves the MNL estimates (compared to Figure 4.1b for the smaller design), but does not seem to improve MNP estimates much. Of course, previous results for MNP did not show much bias, but one would expect to obtain better results from this more complex experiment, especially for the ASC's. On the other hand, more coefficients were estimated due to the alternative-specific utility specifications (18 versus 8 for generic), and the data sets were small, but it is nonetheless surprising that *between* orthogonality did not improve MNP estimates significantly more. MNL and Identity probit exhibit once again a strange pattern for the ASC's for the different alternatives. A graph of these estimates shows that coefficients specific to one alternative (β's) can all be multiplied by a constant to produce those of another alternative. This suggests that non-IID errors impact proportionally all β's of a common alternative, and the proportionality is even more evident in this experiment than the previous one. Consequently, if generic coefficients are expected, and such proportionality between β's of different alternatives is observed, it is likely due to non-IID errors. Significantly, MNP produces more efficient estimates in this case, but it does not seem to retrieve the covariance structure well, as shown by the results in Table 4.6b.

Table 4.6b: Average estimates of the covariance matrix assuming alternative specific coefficients

Labels	True values	Average estimates of the covariance matrix using a 4^{5-3} design	using a 4^{21-18} design
Sigma 11	1 (fixed)	1 (fixed)	1 (fixed)
Sigma 21	-0.5	-0.359	-0.627
Sigma 22	2.5	8.661	4.310
Sigma 31	-1	-1.611	-1.240
Sigma 32	-2.5	-7.771	-4.874
Sigma 33	6	18.066	11.682

Although MNP produces better estimates of both β's and ASC's, it does not estimate the covariance matrix well. The covariance estimates have improved with the larger design, but the true structure cannot be satisfactorily recovered when alternative-specific coefficients are also estimated. As before, if a researcher did not know the true structure, and it was MNP, she nonetheless would make the correct interpretation since all the signs are correct. Once again, the estimates and true values seem to be proportional, behaving as if the structure could be approximated only up to a scale factor. However, in this case the model is theoretically identified, although Keane (1992) notes that the conditions under which it performs well are still unknown; hence investigating different specifications from the data seems to be a good way to evaluate when and what to estimate with MNP models. Finally, even though MNL seems to perform differently for generic and alternative-specific utility specifications, observed biases do not appear to be of consequence on the parameter interpretation; especially since the error structure departed sharply from IID.

4.5 Conclusions

This chapter investigated how IID models, such as MNL and MNP, perform with choice experiments when non-IID errors are present. Two types of designs were considered and generic as well as alternative-specific coefficient utility specifications were examined. Results suggest that in the case of generic coefficients, attribute coefficients (β's) can be estimated efficiently with IID models even in the face of serious departure from IID errors. In the latter case, only the ASC's were affected and biased, which in turn, impacted the model predictions. Trade-offs can be well estimated without the need for the more complex MNP models. However, MNP remains of strategic and practical interest because it actually produced a close approximation of the covariance matrix. This finding was interesting and important since it applied even to designs that were only orthogonal within-alternatives.

For alternative-specific utility specifications, the conclusions were unfortunately different. In this case, non-IID errors affected both the estimated constants and coefficients of all models. MNP estimates exhibited much less bias in the coefficients but could not recover the covariance matrix properly. The between orthogonality property improved estimates of both MNL and Identity probit, but did not improve MNP estimates much; however, the estimated elements of the covariance matrix were closer to the true values for experiments that satisfied this property. These are somewhat disappointing results for MNP, especially as MNP estimates were slightly biased for the experiment which should provide rich estimation data. One needs to ask how well models that assume even more complex error structures and/or utility specifications would perform. At the moment, estimation procedures seem to be ahead of experimental design theory (at least in a choice context) and understanding how well model estimation procedures perform is related to understanding the requirements that need to be satisfied in data. Thus, it would seem that considerably more research such as that conducted in this chapter is needed to define how, when and where one can expect models to perform on particular types of experiments. Extensions of our research could include the investigations of other model forms such as nested logit and mixed logit models, the effect of heterogeneity and estimation of attribute interactions on the choice of these model forms and experimental designs.

Design researchers try to develop better designs that improves the precision of estimated parameters and allow more interesting and complex models such as MNP to be tested. Our results suggest that work remains to be done in this context. On the positive side, however, our results also suggest that under certain circumstances, current experimental design techniques may actually allow one to estimate more complex model structures than previously believed, but unfortunately, this is not true under all conditions.

Chapter 5

Combining Consideration And Choice Data In Hierarchical Models With Covariance Element Selection

Abstract

This chapter presents a new method for carrying out Bayesian inference in a two level hierarchical brand choice model, which allows for a parsimonious representation of both the error and attribute parameter covariance matrices. The size of the covariance matrix increases quadratically with the number of attributes, thus this approach is particularly relevant when dealing with large number of attributes. The first level combines consideration and choice data while the second models consumer heterogeneity. The hierarchical levels are assumed to be normally distributed, which leads to two covariance matrices, and element selection is performed on a decomposition of these matrices to estimate a more parsimonious structure. The assumption of normally distributed latent indirect utilities implies that the first level of the hierarchy is multinomial probit whereas multinomial logit model is used by most of the literature. Estimating a multinomial probit model involves drawing from truncated normal distributions, but makes it easier to exploit a richer preference structure. For example,

one can take advantage of information provided in consideration sets as illustrated in the empirical application in this chapter. Finally, the method provides a way to account for missing information (at random) optimally to model respondents who fail to state preferences in some choice sets. The method is demonstrated empirically by combining choice and consideration information using a Markov Chain Monte Carlo sampling scheme in association with model averaging to estimate model parameters.

5.1 Introduction

Marketers have long been interested in understanding differences between consumers (Smith 1956; Claycamp, Massy 1968; Frank, Massy and Wind 1972; Wind, Yorman 1978; Alderson 1983; Elrod and Keane 1995; Wedel and Kamakura 1998). The ability to assess individual preferences is of primary interest to develop, target and promote more appropriate products that (hopefully) also will be more successful. For those reasons, various methods have been developed to model consumer heterogeneity. Early choice models like the Guadagni and Little (1983) multinomial logit (MNL) model attempted to capture heterogeneity in the population by including demographics as independent variables. These types of model specifications parameterise the alternative specific constants (ASC's) as a function of individual characteristics; and including interactions between demographics and other covariates (such as price) allows one to estimate models that not only provide individual predictions but also individual level parameters. However, such models heavily depend on the quality and relevance of the demographic or psychographic information that only partially account for heterogeneity but do not model it per se.

One way to address this issue is to estimate a set of coefficients for each individual (a fixed effects model), but there rarely is sufficient data available to allow one to estimate such models. Another approach is to estimate only the distribution of the heterogeneity; this distribution is usually referred to as a mixing distribution, which was approximated initially by using a discrete number of mass points (e.g., Kamakura and Russell 1989). A more general approach is to estimate the moments of an assumed continuous, joint multivariate distribution of heterogeneity (e.g., Chintagunta, Jain and Vilcassim 1991). Specifically, conditional on each respondent's parameters, one can assume the model to be logit or probit, but individual parameters are unknown and so must be integrated out of the model by making an assumption about their distribution.

The Kamakura and Russell (1989) mixture regression model approach is fast and easy to estimate, and is compatible with the concept of market segmentation. However, one computational issue is the fact that multiple optima may be obtained; hence, to be confident that the global and not a local optimum solution is attained, models need to be estimated a number of times with different starting values. The latter is a minor limitation of the approach because the method produces results quickly. Mixture regression is a discrete approach to heterogeneity because it assumes that each segment is a homogeneous sub-population (Wedel and Kamakura 1998 reviewed these methods). Such segments are very appealing managerially, which probably is one reason why they have been widely applied in marketing. The main advantage of uncovering segments is that it allows firms to introduce product lines customised to specific groups of consumers. Each segment may be targeted with either a variant of the main product or a specific marketing program. However, despite its obvious managerial and behavioural appeal, this modeling approach does not completely account for heterogeneity because it is a discrete approximation to a potentially continuous problem.

Thus, the reality of homogeneous segments can be questioned, which suggests that one should consider approaches that can estimate continuous distributions of parameters (e.g., Allenby, Arora and Ginter 1998; Allenby and Rossi 1999). Such approaches are variously termed "random coefficients," "mixed logit" (or probit) and/or "hierarchical logit" (or probit), depending on the error assumptions and computational methods involved. The terms "random coefficients," "mixed or heterogeneous logit" models are typically used by frequentists, whereas Bayesians refer to them as "hierarchical models". The objective of both approaches is to estimate moments such as means and variances of heterogeneity distributions (sometimes called "hyper-parameters"). The assumption of a continuous distribution has the advantage of being able to account for a perfectly heterogenous population; hence extreme behaviours can be better represented than with mixture models (Allenby and Ginter 1995). Another advantage is that individual level parameters can be obtained (See Rossi and Allenby, 1993 for a pure Bayesian version; and Revelt and Train, 1999 for a maximum likelihood version). However, a potential drawback is that one has to make a distributional assumption, which is most frequently the multivariate normal (but frequentists also use the lognormal – see, e.g. McFadden and Train, 2000). However, it is worth noting that mixture models can be viewed as a non-parametric approach because they do not require a priori distributional assumptions.

The foregoing suggests that continuous distributions should produce more precise market simulators than discrete approximations (Allenby, Arora and Ginter 1998), although many companies cannot afford to customise products for each customer. Firms that cannot personalise products should benefit less from individual parameters, and instead would have to find optimal numbers of product variants or develop marketing programs based on: (1) the heterogeneity in the population, and (2) the costs and benefits of targeting a specific number of segments. This suggests that such firms would benefit from discretisation of heterogeneity; which mixture models achieve directly (but to date, none have been proposed that also take profitability into account). Thus, like most modeling methods, discrete and continuous approaches to model heterogeneity each have their advantages and disadvantages.

The model proposed in this chapter assumes a continuous distribution of heterogeneity. As noted above, both Bayesian and frequentist methods may be used to estimate such models. On the one hand, classical methods must integrate over the individual parameters and simulate the probabilities in order to evaluate and maximise the likelihood. On the other hand, Bayesian techniques need to draw from the conditional distribution of all parameters; and depending on the error assumptions, this distribution may be unknown and may require the Metropolis-Hastings (MH) algorithm (Allenby and Ginter 1995). Both approaches seem to compare favourably with one another, so as noted by Revelt and Train (1999), choice of method is mostly a case of which one makes a researcher feel more comfortable.

The method proposed in this chapter is Bayesian. It addresses the problem of parsimony, which is specific to the assumption of a continuous distribution and becomes increasingly relevant as numbers of variables increase. Effectively, a multivariate normal distribution involves the estimation of a covariance matrix, which contains $K(K+1)/2$ parameters (K = the number of covariates in the model); that is, as the number of variables reaches 30, 465 elements have to be estimated.

To address this issue, we propose a selection procedure, the objective of which is to reduce the number of estimates to obtain a more parsimonious representation of heterogeneity. For example, one way to reduce the number of elements is to estimate a diagonal covariance matrix, which might be considered adequate from a statistical point of view. However, if segments exist (homogeneous or not) this is likely to lead to a non-zero correlation structure; an example of this is a market that exhibits two distinct groups of customers; one highly sensitive to price and insensitive to service and the other exactly the reverse. In the latter case, if one models heterogeneity as a continuous distribution, one should uncover a correlation between price and service because as customers' sensitivity to price decreases, their sensitivity to service increases. This implies that there would be non-zero off diagonal elements in the covariance matrix and a diagonal structure would represent a constraint. Thus, consistent with this line of reasoning, the main contribution of this research is the application of a selection procedure to model heterogeneity by selecting and estimating the relevant off-diagonal elements of the covariance matrix.

In designed choice experiments, respondents typically are asked to evaluate a set of options and choose the one they prefer or the one they would buy. Given this choice information and assuming that Random Utility Theory applies, one can estimate the respondents' indirect utilities (e.g., by estimating MNL or MNP models). An advantage of this choice elicitation procedure is that it minimises respondents' cognitive burden, and it simulates the types of decisions that they make in everyday life. Unfortunately, however, observing only a respondent's most preferred option ignores the potential for obtaining additional information from other responses. For example, traditional conjoint experiments typically ask respondents to rate options on category rating scales, which would seem to provide more in-depth information than discrete choices. However, the quality and validity of such scales have been questioned (e.g. Louviere, Hensher and Swait 2000) because of the assumptions made about human cognitive abilities to produce metric data in a reliable and valid manner.

The latter point also was made by Chapman and Staelin (1982) among others, who asked their respondents to rank all options. Ranking responses require less restrictive assumptions about human ability than ratings because one needs only assume ordinality of responses, and rankings provide more information than discrete choices. Moreover, Beggs, Cardel and Hausman (1981) demonstrated that if the errors were distributed as extreme value type I random variates, the indirect utility distribution associated with the preferred option is independent of the ranking of the other options. Thus, the probability of a specific ranking of options can be written as a product of conditional probabilities, which allows one to exploit rankings to obtain a ranking utility model. For example, suppose that three alternatives a, b, c were ranked as follows: $a \succ b \succ c$ (where $\succ$

means preferred to); hence, the probability $p(a \succ b \succ c)$ equals the product of the following multinomial logit probabilities (assuming extreme value type I errors):

$$p(a \succ b \succ c) = p(a \succ b, a \succ c)\, p(b \succ c),$$

where $p(a \succ b \succ c)$ is the probability that a is the preferred option and b is second best. Chapman and Staelin (1982) also used this property to exploit rank ordered choice sets; more recently, Layton and Lee (1998) extended their method to account for ties, which allows ratings to be treated as ordinal measures rather than interval scales.

It is worth noting that the Beggs et al.'s (1981) result does not imply that other distributions like the normal will not produce a random utility model. Their result only implies that assuming extreme value type I distributed errors makes it easy to compute the probability of a ranked choice set. Indeed, they noted that the normal also could be considered, but evaluation of the probabilities would be computationally challenging because the choice probabilities do not have a closed form expression. Hajivassiliou and Ruud (1994) demonstrated how to estimate such a model with standard approximation methods when the number of alternatives is small ($J \leq 4$), or by using the GHK simulator for larger numbers of alternatives (e.g., see Geweke 1991; Hajivassiliou and Ruud 1994; Keane 1994). In a related vein, Hajivassiliou and Ruud (1994, p. 2410) noted that: "The rank ordering yields considerably more information about the underlying preference parameters than the simpler, highest-ranked-alternative response. Hence, consumer survey designers often prefer to ask for complete rankings."

The preceding discussion suggests that obtaining other preference information in addition to discrete choices should prove valuable in choice modeling. Consequently, the Bayesian modeling approach proposed later in this chapter illustrates how additional information (e.g., a complete ranking) can be incorporated in a hierarchical model. Specifically, by assuming normally distributed indirect utilities, Bayesian methods can easily deal with additional preference response data. Of course, the previous discussion suggests that additional preference information can be obtained by observing a complete ranking; however, this approach is not without limitations: (1) ranking tasks increase in difficulty as numbers of alternatives increase; (2) each rank increases parameter estimation efficiency, but at a decreasing rate, as noted by Chapman and Staelin (1982), who suggested that an explosion depth of more than three probably would not be beneficial. Similarly, Ben-Akiva, Morrikawa, Shiroishi (1991) found that response reliability decreased with increasing rank.

Swait and Louviere (1993) proposed another way to obtain additional preference information by exploiting the extra information in consideration response data with standard MNL models. Compared with rankings, consideration elicitation procedures have the advantage that they can be used with larger choice sets and still provide valuable information. Considered options may be assumed to be preferred to non-considered ones, which information can be used to obtain more precise parameter estimates. For example, section 5.4 discusses an application in which subjects were asked to reveal not only the option they like best but also those they would consider. The model and estimation procedure proposed use this extra information to estimate individual preferences by selecting a parsimonious covariance structure and accounting for non-IID errors in the indirect utilities. The selection process is data driven and allows the model to converge to an IID structure with a diagonal covariance matrix to model heterogeneity if supported.

The rest of the chapter is organised as follows. Section 5.2 reviews hierarchical models in more detail. Section 5.3 presents the sampling scheme of the proposed method, which is empirically tested in Section 5.4. Conclusions end the chapter.

5.2 Hierarchical models

This section discusses two-level hierarchical models in more detail and introduces the selection procedure proposed in this research. It should be noted that hierarchical models are well-known in the Bayesian literature and have been applied in conjunction with continuous dependent variables (Lindley and Smith 1972) as well as with discrete dependent variables using data augmentation (e.g. Albert and Chib 1993) or the Metropolis Hasting algorithm (e.g. Allenby and Ginter 1995). The proposed model is typical of many marketing applications in that the first level of the hierarchy models individual preferences as a function of covariates. This level assumes a normally distributed stochastic component, and can estimate a full covariance matrix (Σ) equivalent to MNP.

The second level of the hierarchy, which models individual level taste parameters as a function of variables that characterise respondents, also assumes normally distributed errors, which in turn introduces a second variance covariance matrix (Ω). This second covariance matrix provides information about differences in individual parameters; for example, large variances suggest that preferences for a particular attribute may differ considerably from one individual to another, whereas a small variance demonstrates more agreement between individuals. In the latter case, the population can be said to be homogeneous with respect to that attribute.

Non-zero covariances reveal patterns of differences in individual preferences; for example, respondents who are less sensitive to price may be more sensitive to quality and vice versa, which would lead to non-zero covariance elements. Although non-zero elements seem likely, it also may be that some should be zero. Moreover, even if the true model has a full covariance structure, the large number of parameters involved and the limited amount of information in choice data would make it difficult to obtain precise estimates (Keane 1992 discusses "parameter fragility" in MNP models).

Thus, Elrod and Keane (1995) proposed a factor-analytic probit structure that reduces the number of estimated elements. For similar reasons, the objective of the proposed approach also is to reduce the number of elements to be estimated. However, the Bayesian selection procedure used should achieve better resolution because it does not impose a specific number of factors on the covariance matrix. A small number of elements is estimated only if the data support this simplification. When a full structure should be employed the procedure aims to select all elements and estimate the complete variance covariance matrix. Also if the data are not rich enough to reveal the importance of some elements the method should simply not select them.

The proposed model is as follows:

$$U_{ij} = x_{ij}\beta_i + \varepsilon_{ij}$$

$$\beta_i = \Gamma z_i + \mu_i$$

where $i = 1$ to N indexes individuals, $j = 1$ to J indexes alternatives,

U_{ij} represents a latent indirect utility for alternative j associated with individual i,

x_{ij} is a vector of K covariates, which describe alternative j for individual i,

β_i is a vector of K coefficients for individual i,

ε_{ij} represents a normally distributed error $N(0,\Sigma)$,

z_i is a vector of L covariates characterising individual i,

Γ is a $(K \times L)$ matrix of coefficients relating β_i with z_i,

μ_i represents an error assumed to be normally distributed $N(0,\Omega)$.

From a modeling perspective, the first difficulty is that, although subject's choices are known, the indirect utilities U_{ij} remain unobserved. However, in a random utility model, respondents are assumed to select the option that maximises their utility and y_{ij} can therefore be defined as follows:

$$y_{ij} = 1 \text{ if } U_{ij} > U_{ik} \text{ for any } k \neq j$$

$$y_{ij} = 0 \text{ otherwise.}$$

Random Utility Theory was first proposed by Thurstone (1927) to explain dominance judgements between pairs of options; that is, consumers should try to choose the option they prefer, subject to constraints such as price and income. Standard economic theory supposes that goods provide individuals with some value referred to as *utility,* which is a function of a basket of goods. Any basket of goods can be associated with a specific utility, which is further defined as a *direct utility*. Consumers are assumed to trade-off the direct utility of a good (or a basket of goods) with its price, accounting for their budget constraint (income). The consumers' objective is to make decisions that maximise their utility. Typically, consumers are assumed to reach different maxima depending on the pricing structure and their level of income. The equation characterising all maxima (as a function of price and income) is termed the *indirect utility function*; that is, indirect utilities are the solutions of an optimisation process undertaken by each consumer. From the researcher's position, a consumer may not choose what looks like the best option; however, this discrepancy between researchers and respondents can be accounted for by adding a random component to the utility function. Thus, consumers' utilities consist of a systematic component $x_{ij}\beta_i$, known to the researcher, and a random component ε_{ij}, known to the respondent only. The justification for the random component is that the researcher may not specify the model accurately (omitted variables) or may err in measuring the responses.

As noted earlier, frequentists call the above model an MNP model with heterogeneity (e.g., McFadden and Train 2000), and estimate the full covariance structure discussed or assume a simplified error structure (for example, assuming a diagonal covariance matrix). In the next section, we present the sampling scheme used to estimate this model with Markov Chain Monte Carlo methods assuming a data driven selection process on the elements of the covariance matrix to obtain a more parsimonious model.

5.3 Sampling Scheme

The standard sampling scheme corresponding to the model proposed in section 5.2 involves drawing from the following distributions:

1. $\pi\left(U_i \mid \beta_i, \Sigma\right)$, where U_i is a vector containing all U_{ij} for respondent i.
2. $\pi\left(\beta_i \mid U_i, \Sigma, \Gamma, \Omega\right)$
3. $\pi\left(\Gamma \mid \beta_i, \Omega\right)$
4. $\pi\left(\Sigma \mid U_i, \beta_i\right)$
5. $\pi\left(\Omega \mid \beta_i, \Gamma\right)$

However, the approach proposed in this chapter also considers selecting elements of both covariance matrices. Thus, the full sampling scheme includes additional binary latent variables that specify more or less complex covariance matrices at any given iteration. This is discussed later in this section with the sampling of the covariance matrices Σ and Ω. We now consider in more detail the sampling of the distributions in steps 1. to 5. above.

The first distribution $U_i \mid \beta_i, \Sigma$ is a truncated multivariate normal. This distribution is very interesting because the observed discrete outcome (we only observe which option has been selected) has an underlying continuous latent distribution. This draw corresponds to a data augmentation (i.e. Tanner and Wong 1987) approach proposed by Albert and Chib (1993) to simplify the analysis of discrete response models (see also

Chapter 3). They show how augmenting the problem by drawing what is referred to in this chapter as the (indirect) utilities, simplifies the evaluation of all other conditional distributions. In a binary setting, drawing the utilities is straightforward in that the utilities associated with the preferred option can be drawn from a normal distribution truncated above zero with the other utilities truncated below zero.

Unfortunately, the model presented in this chapter involves more than a binary selection, and multinomial responses present an additional difficulty because the truncated distribution becomes multivariate. However, a draw from this distribution can be obtained by sampling from a series of univariate distributions (McCulloch and Rossi 1994; Hajivassiliou and Ruud 1994; Hajivassiliou and McFadden 1998). The idea is to draw each utility U_{ik} conditional on all others (U_{ij} with $j \neq k$); in that way the utility corresponding to a selected alternative can be drawn from a normal distribution truncated above the maximum of all non-selected utilities. Similarly, for non-selected alternatives, utilities are truncated below the utility of the selected option.

If more information is known, such as a complete ranking, this procedure does not become more complex. For example, the utility of the first ranked option is truncated above the second ranked option, which in turn is truncated below the first but above the third ranked option. This process continues until a utility has been drawn for each ranked option. Drawing in turn from these univariate distributions yields a Gibbs sampler (Casella and George 1992 discuss the Gibbs sampler in detail) that converges to the truncated multivariate distribution. Note that if two alternatives are ranked first, each utility is truncated above the utility of the next ranked option.

The above example used a rank ordering of options to facilitate exposition of the process, but the utilities corresponding to any sort of rank information can be sampled in the same way. In the empirical application considered in section 5.4, the utilities are drawn so as to respect the order implied by the consideration data and the preferred choice; that is, the utilities of all considered options must be larger than the utilities of non-considered options and the utility of the preferred option must be the largest.

The second distribution can be sampled as follows:

$$\pi(\beta_i \mid U_i, \Sigma, \Gamma, \Omega) \propto \pi(U_i \mid \beta_i, \Sigma, \Gamma, \Omega)\,\pi(\beta_i \mid \Gamma, \Omega)$$

$$\pi(\beta_i \mid U_i, \Sigma, \Gamma, \Omega) \propto \exp\left(-\frac{1}{2}\sum_l (U_{il} - X_{il}\beta_i)' \Sigma^{-1} (U_{il} - X_{il}\beta_i)\right)\pi(\beta_i \mid \Gamma, \Omega)$$

where the summation is over l,

l indexes each observation (choice sets) for respondent i and

$$\pi(\beta_i \mid \Gamma, \Omega) \propto \exp\left(-\frac{1}{2}(\beta_i - \Gamma z_i)' \Omega^{-1} (\beta_i - \Gamma z_i)\right).$$

Thus,

$$\pi(\beta_i \mid U_i, \Sigma, \Gamma, \Omega) \propto \exp\left(-\frac{1}{2}(\beta_i - \bar{\beta}_i) V_{\beta i}^{-1} (\beta_i - \bar{\beta}_i)\right)$$

where $\bar{\beta}_i = V_{\beta i}\left(\sum_l X_{il}'\Sigma^{-1}U_{il} + \Omega^{-1}\Gamma z_i\right)$ and $V_{\beta i}^{-1} = \sum_l X_{il}'\Sigma^{-1}X_{il} + \Omega^{-1}$.

As a result, $\beta_i \mid U_{ij}, \Sigma, \Gamma, \Omega$ is normally distributed with mean $\bar{\beta}_i$ and variance $V_{\beta i}$.

The third distribution to be sampled is another multivariate normal.

$$\pi(\Gamma \mid \beta_i, \Omega) \propto \exp\left(-\frac{1}{2}\sum_i (\beta_i - \Gamma z_i)' \, \Omega^{-1} (\beta_i - \Gamma z_i)\right)\pi(\Gamma).$$

Using the following property $vec(Ab) = (b' \otimes I)vec(A)$, where A is a matrix and b a vector of appropriate dimension, we can write:

$$\pi(\Gamma \mid \beta_i, \Omega) \propto \exp\left(\frac{-1}{2}\sum_i (\beta_i - (z_i' \otimes I)vec(\Gamma))' \Omega^{-1} \sum_i (\beta_i - (z_i' \otimes I)vec(\Gamma))\right)\pi(\Gamma).$$

We define $\gamma = vec(\Gamma)$, and have

$$\pi(\gamma \mid \beta_i, \Omega) \propto \exp\left(-\frac{1}{2}(\gamma - \overline{\gamma})' V_\gamma^{-1} (\gamma - \overline{\gamma})\right)\pi(\gamma)$$

where $\overline{\gamma} = V_\gamma \sum_i (z_i' \otimes I)' \, \Omega^{-1} \beta_i$

$$V_\gamma^{-1} = \sum_i (z_i' \otimes I)' \, \Omega^{-1} (z_i' \otimes I) = \sum_i z_i z_i' \otimes \Omega^{-1}.$$ [1]

Therefore, if a normal distribution is assumed for γ such as $N\left(\gamma^0, V_{\gamma^0}^{-1}\right)$ then $\gamma \mid \beta_i, \Omega$ can be sampled from a normal distribution with mean $\left(V_\gamma + V_{\gamma^0}\right)^{-1}\left(\overline{\gamma} + V_{\gamma^0}\gamma^0\right)$ and variance $V_\gamma + V_{\gamma^0}$.

We now turn to the fourth and fifth distributions. To sample Σ and Ω we use a decomposition and sample on the decomposed elements. That is, we set $\Sigma^{-1} = B'DB$

and $\Omega^{-1} = G'HG$, where B and G are lower triangular matrices with all diagonal elements equal to 1, while D and H are diagonal matrices. Smith and Kohn (1999) used this decomposition of the inverse of a covariance matrix to estimate seemingly unrelated regression models and the method was applied to MNP models in Chapter 3 to impose the identification restrictions that such models require. It should be noted that this model requires further identification restrictions, but for the moment these are ignored to simplify the exposition, because the method proposed for sampling the conditional distributions of B and D can be applied readily to G and H, respectively.

We first consider drawing the diagonal elements $d_1,\ldots,d_J$ of D. Given the individual parameters β_i the likelihood is:

$$\pi\left(U \mid \beta_i, D, B\right) \propto |B'DB|^{N_J/2} \exp\left(-\frac{1}{2}\sum_j \sum_i \varepsilon_{ij}' B'DB\varepsilon_{ij}\right) \tag{5.1}$$

thus,

$$\pi\left(U \mid \beta_i, D, B\right) \propto \prod_i d_i^{N_J/2} \exp\left(-\frac{1}{2} trBAB'D\right), \tag{5.2}$$

$$\pi\left(d_i \mid \beta_i, U, B\right) \propto \prod_i d_i^{N_J/2} \exp\left(-\frac{1}{2} trBAB'D\right) p\left(d_i \mid B\right)$$

because $\det(B) = 1$. In (5.2), the matrix $A = \sum_j \sum_i \varepsilon_{ij}\varepsilon_{ij}'$, U is a vector containing all U_{ij}'s and N_J represents the total number of choice sets.

---

[1] Because $\left(z_i' \otimes I\right)' \Sigma^{-1} \left(z_i' \otimes I\right) = \left(z_i' \otimes I\right)' \left(1 \otimes \Sigma^{-1}\right)\left(z_i' \otimes I\right)$ and $\left(A \otimes B\right)\left(C \otimes D\right) = AC \otimes BD$ if AC and BD exist (Magnus and Neudecker 1988 page 28).

Therefore, up to a prior $p(d_i \mid B)$, $d_1,...,d_J$ have independent gamma distributions. If the prior density of d_i is $\text{Gamma}\left(d_i^0/c_{di}, 1/c_{di}\right)$, then $\pi\left(d_i \mid U, \beta_i, B\right)$ is $\text{Gamma}\left(N_J/2+1+d_i^0/c_{di}, C_{ii}/2+1/c_{di}\right)$ where $C_{ii}=BAB'$ and c_{di} is a positive scalar. This choice of prior is motivated by the result in Chapter 3 suggesting that the prior variance of the d_i elements needs to be moderately tight in MNP models. Although simulations suggest that this has little impact on the posterior distribution of the parameters with well identified data, it is theoretically the case that if the mean prior is not particularly well centered (with the posterior mean of the likelihood), a tight prior variance will impact the posterior estimates.

Diffuse priors typically reflect ignorance of the posteriors a priori and ensure that the mean priors have no bearing on the posterior means. For mean priors to have little impact but still impose a moderately tight prior variance, the priors are updated or re-centered after a few iterations. This approach has the advantage that it can be easily implemented and works well in simulations (not reported here). Nevertheless, the prior variances should not be too tight, otherwise it will be reflected in the posterior variances. In summary, this extra step may slow down the convergence process when starting from badly chosen mean priors, but once the latter are updated and centered, it ensures that the priors have no impact on the posteriors.

The prior mean equals d_i^0 (i.e., $d_i^0 = d_i^0/c_{di}/1/c_{di}$) and the prior variance equals $c_{di}d_i^0$ (i.e., $c_{di}d_i^0 = d_i^0/c_{di}/1/c_{di}^2$). Therefore, setting $c_{di}=5$, for example, reflects a prior variance 5 times as large as the mean prior, and so once convergence is achieved this choice of prior has little impact.

To complete the sampling of the covariance matrix we need to consider the conditional distribution of B. Following Chapter 3, we partition the matrices B and A as follows:

$$B=\begin{bmatrix}1 & 0 & 0 & 0\\ b_2' & 1 & 0 & 0\\ \vdots & & 1 & 0\\ b_J' & & & 1\end{bmatrix} \qquad A=\begin{bmatrix}A_i & a_i & .\\ a_i' & \alpha_i & .\\ . & . & .\end{bmatrix}$$

where b_i are vectors $(i-1)\times 1$, A_i and a_i are dimensioned conformably as $(i-1)\times(i-1)$ and $(i-1)\times 1$. Using this notation, we have from (5.2),

$$\pi(b_1,\ldots,b_J \mid U,D,\beta)\propto\prod_{i=1}^{J}\exp\left(-\frac{1}{2}d_i(b_i'A_ib_i+2b_ia_i)\right)\pi(B\mid D).$$

It is clear now that up to the prior $\pi(B\mid D)$, the vectors $b_1,\ldots,b_J$ are independent and Gaussian. Hence, if the b_i are normal and independent a priori with $b_i\sim N\left(b_i^0,V_{bi^0}^{-1}\right)$, then $\pi(b_i\mid U,\beta,D)$ is $N\left(\overline{b}_i,V_{bi^0}^{-1}\right)$, with $V_{bi}^{-1}=A_id_i+V_{bi^0}$ and $\overline{b}_i=-V_{bi}^{-1}\left(a_id_i-V_{bi^0}b_i^0\right)$.

It is straightforward to find similar distributions for G and H because the conditional distribution of β_i given Γ, G and H,

$$\pi\left(\beta_i\mid\Gamma,H,G\right)\propto\mid G'HG\mid^{N_i/2}\exp\left(-\frac{1}{2}\sum_i\mu_iG'HG\mu_i\right),$$

is an expression equivalent to (5.1).

Contrary to Σ, Ω can be sampled without constraint. In the case of Σ, there are two identification issues because the U_{ij}'s are unobserved. Specifically, a linear transformation (with a positive slope) of the utilities leaves the ordering, and thus the choices, unaffected and produces an identical model specification. The model can be identified by fixing one alternative (referred to as the base) to ensure that adding a constant will no longer preserve the ordering. Furthermore, the variance of another alternative needs to be fixed to ensure that multiplying the utilities by a positive constant leads to a different model. The sampling can easily accommodate both restrictions by setting $d_1 = 1$ and $b_J = 0$.

If a diagonal covariance matrix is estimated, which simply means setting all b_i to zero, d_J can be estimated but cannot be if one samples a full covariance matrix (see Bunch 1991). Therefore, in accord with Chapter 3, we first estimate a diagonal structure to obtain an estimate of d_J, and set d_J to this estimate to see if more covariance elements should be selected, using an element selection procedure. Consequently, if the data do not support any covariance elements (i.e. they all equal zero), the approach should retrieve this structure and set all elements of the B matrix to zero.

Sampling Ω is facilitated because no identification issue arises. Note that if only one choice set were administered to each respondent this would not be true because Σ would be perfectly confounded with the first rows and columns of Ω. In such a case Σ should be set to the identity and $\Omega_{11} = 1$. Typical choice experiments obtain responses to multiple choice sets, so the restriction does not apply but Σ can be expected to be weakly identified if the number of choice sets administered to each respondent is low.

Distributions 4. and 5. introduce two indicator variables λ_{ij} and η_{ij} that indicate the elements of B and G that should be selected and estimated. The selection mechanism operates by calculating the probability that each element of B and G differs from zero. At each iteration, the values of the indicators are drawn according to the probability of their associated element being zero. The states of the binary variables then specify the elements of B and G to estimate. B_λ and G_η are now indexed by λ_{ij} and η_{ij} to reflect this dependence.

To compute the probability that an element of B_λ equals zero, we set the following prior density for $b_{i\lambda_i} \mid D, \lambda_{ij}$ to $N\left(b^0_{i\lambda}, c_{bi}\left(A_{i\lambda} d_i\right)^{-1}\right)$, where c_{bi} is a positive scalar, which makes the prior tighter (or looser) around the mean prior.

By integrating $b_{i\lambda_i}$ out, we have:

$$\pi\left(\lambda_i \mid U, \beta_i\right) \propto d_i^{(N_J - q_{\lambda_i})/2} r^{q_{\lambda_i}/2} \mid A_{i\lambda_i} \mid^{-1/2} \exp\left(-1/2\left(\alpha_i - r a'_{i\lambda_i} A_{i\lambda_i}{}^{-1} a_{i\lambda_i} - a'_{i\lambda_i} b^0_{i\lambda_i} r_1\right) d_i\right)$$

where $r = c_{bi}/(1+c_{bi})$, $r_1 = 1/(1+c_{bi})$ and

q_{λ_i} represents the number of selected elements in $b_{i\lambda_i}$.

Thus, $\pi(\lambda_i \mid U, \beta_i)$ is gamma distributed in d_i and this expression can be used to draw d_i unconditional on $b_{i\lambda_i}$, which may improve the convergence properties of the sampler in some cases.

It is now possible to integrate out d_i as a gamma integral:

$$\pi(\lambda_i \mid U, \beta_i) \propto \Gamma(\varphi) * r^{q_{\lambda_i}/2} \mid A_{i\lambda_i} \mid^{-1/2} \left(1/2(\alpha_i - ra'_{i\lambda_i} A_{i\lambda_i}{}^{-1} a_{i\lambda_i} - a'_{i\lambda_i} b^0_{i\lambda_i} r_1) + 1/c_{di}\right)^{-\varphi}$$

where $\varphi = (N_J - q_{\lambda_i})/2 + d_i^0/c_{di}$.

This probability does not depend on Σ and can be evaluated up to a constant of proportionality for both states of the binary variable λ_i (i.e., the element $b_{i\lambda_i}$ included and excluded from the sampling scheme), and computed exactly via normalization (their sum must equal 1). An identical development is obtained for the elements of G_η, which completes the sampling scheme.

5.4 Empirical Illustration

The application examined in the present research considers different offerings of financial services; more specifically, the objective is to investigate consumer preference trade-offs for various services related to a particular type of account. The data used in the empirical example were collected as part of a larger research project funded by a grant from NCR's Knowledge Lab to Professor Jordan Louviere at the University of Sydney. In accordance with agreements with the Knowledge Lab, the actual nature of the product category is disguised, but can be revealed to be a type of financial services account. The attributes and levels considered as well as the experimental design are for the most part hidden but a discussion of the choice task is provided in the next paragraphs, after which the estimation results follow.

The Choice Experiment

The choice experiment manipulated 26 attributes, 17 with 4 levels and 9 with 2 levels. Among the latter 9 attributes, 5 applied only to large and moderately-large institutions. A more complete description of the attributes and levels is provided in Table 5.1; rows represent attributes that describe financial institutions and questions asked in each choice set. Columns depict types of institutions (e.g., large, moderately-large, small and non-traditional). Table 5.1 shows the number of levels used for each attribute; the 2-level attributes are lightly shaded. Prime interest rate varied over 4 levels, and all interest rates shown to respondents were deviations from the prime. A new account feature that was not available at the time but that potentially would allow more convenient access to account services also was included; it is listed in the table as

"Convenience Feature". Additionally, four brands of institutions were varied under each institution type, but version of the survey presented only one brand under each type.

The choice experiment was designed using the method discussed by Louviere and Woodworth (1983) in which the attributes of all choice alternatives are treated as a collective factorial and a main effects design is selected in such a way that attribute columns are orthogonal within- and between-alternatives. As previously indicated, this design can therefore be expressed a $16 \times 4^{65} \times 2^{26}$ factorial; the main effects design was constructed from a 16^{17-15} fractional factorial that generates 256 choice sets. Basing the design on 16-level columns allowed it to be blocked into 16 versions of 16 choice sets. Most 16-level columns were transformed into orthogonal main effects designs based on the 4^{5-3} factorial; some of which were in turn transformed into three 2 level columns (see Dey and Mukerjee 1999). This design process is illustrated in Figure 5.1.

Figure 5.1: Use of the 17 attribute columns with 16 levels

	16 levels	**4 levels**	**2 levels**	**Used**
16 levels	17			1
4 levels	→	80		65
2 levels		→	45	26

A balanced incomplete block design (BIBD) also was used to assign four brands to each institution type in each of the 16 versions.

Table 5.1 Description of choice set from the financial study

Type of account (hidden) Scenario 1,...,16 Version # 1,...,16 *Prime Interest rate:PI1 %, PI2 %, PI3 %, PI4 %*	Large, international institution	Moderate-large regional institution	Small/local institution	Non-traditional institution
Name of the financial institution	**1,2,3,4**	**1,2,3,4**	**1,2,3,4**	**1,2,3,4**
Hidden attribute	$A	$B	$C	$D
Account Type A				
Monthly account fee	$A	$B	$C	$D
Fee to take money out	$A	$B	$C	$D
Fee to put money in	$A	$B	$C	$D
Annual interest paid on account balance	Prime -A%	Prime –B%	Prime –C%	Prime -D%
Account Type B				
Monthly account fee	$A	$B	$C	$D
Fee to take money out	$A	$B	$C	$D
Fee to put money in	$A	$B	$C	$D
Annual interest paid on account balance	Prime -A%	Prime –B%	Prime –C%	Prime -D%
Transferring between accounts				
Fee to transfer to another account in your own institution	$A	$B	$A	$B
Fee to transfer to an account in another institution	$A	$B	$A	$B
Line of credit				
Fee to include a line of credit	$A	$B	$A	$B
Link to credit card (CC)				
Annual fee	$A	$B	$C	$D
Annual interest	Prime -A%	Prime –B%	Prime –C%	Prime -D%
Loyalty scheme – associated with airlines	None	Airline 1	Airline 2	Airline 3
Internet access (IA)				
Fee to use Internet access for services type 1	$A	$B	$C	$D
Fee to use Internet access for services type 2	$A	$B	$C	$D
Telephone access (TA)				
Fee to use telephone access for services type 1	$A	$B	$C	$D
Fee to use telephone access for services type 2	$A	$B	$C	$D
Convenience features (CF)	Level 1	Level 2	N/A	N/A
Attribute (hidden)	Level 1	Level 2	N/A	N/A
Attribute (hidden)	Level 1	Level 2	N/A	N/A
Locations where you can deal in person (LP)				
Nearest full-service office	Level 1 or 2	Level 1 or 2	Level 1 or 2	Level 1 or 2
Retailer type 1	Retailer 1	Retailer 2	N/A	N/A
Retailer type 2	Retailer 1	Retailer 2	N/A	N/A
Which accounts would you seriously consider if you had to switch today for whatever reason? ☑ **AS MANY** boxes as apply.	☐	☐	☐	☐
If you had to switch your account today for whatever reason and the five options on the right were your only options, which would you choose? ☑ **ONLY ONE** box.	☐	☐	☐	☐
	☐ I'd use X			
If your current account or the four accounts on the right were your only options, which of the following would you most likely choose. ☑ **ONLY ONE** box.	☐	☐	☐	☐
	☐ I'd keep my current account & not switch			

Table 5.1 shows that respondents were asked to state the following in each choice set (1) which institutions(s) they would seriously consider, (2) their preferred option among the four institutions presented or a constant option and, (3) their preferred option among the four offered institutions and their current account.

Surveys were administered to a sample of 370 students enrolled in an American university. Each student was randomly assigned to one of the 16 survey versions. It is worth noting that an advantage of asking more than one question in each choice set is that it allows one to check response validity. For example, respondents who state they would switch from their current account to one of the proposed options should not choose the "I'd use X" option, and should state that they would consider the chosen institution in the consideration question. Such inconsistencies, especially if repeated in most choice sets, signal respondents who do not take the task seriously or do not understand the questions. These checks were performed on the data set to remove unreliable responses; in addition, some surveys were discarded due to coding errors. The final data set used in the analysis below contained responses from 350 respondents.

It also is worth noting at this stage that this preference elicitation approach ensures that more information is gathered for each choice set. Specifically, in the present case around 55% of the responses to question 3 were "I'd keep my current account" which means that, had only this question been asked, there would have been considerably less information about trade-offs. That is, choice experiments that include constant choice options run the risk that respondents will select them in each choice set, thereby providing no information about preferences. The latter is an extreme case, but including a second question with a less attractive constant option ensures that enough information is gathered to still estimate trade-offs for respondents who truly prefer their current option. Thus, one needs to be careful and pilot test the attractiveness of such constant options; for example in the present survey the option "I'd use X" was selected less than 5% of the time, which guarantees that one can obtain trade-off estimates.

Results

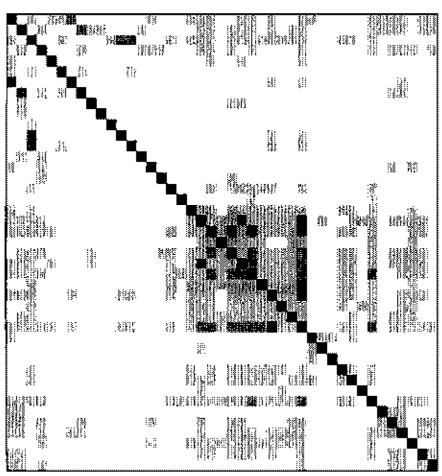
The model assumes the following values for the prior distributions obtained in section 5.3: $c_{bi} = c_{di} = 1$, $c_{gi} = c_{hi} = 10$, $\gamma = N(0,10I)$ and $d_i^0 = 1$, thus the priors for the first level of the hierarchy correspond to an IID probit model. Setting c_{bi} and c_{di} to very small values (e.g. 10^{-10}) allows estimation of an IID structure without any change to the current sampling scheme. Tight priors represent intermediate models between an IID and a full covariance structure that can be useful when the model parameters are weakly identified. No other specific prior information needs to be incorporated in the analysis so the mean priors were selected to have minimal impact on the posterior means. Usually, large prior variances would be assumed in order to reflect such a lack of a priori knowledge. However, MNP models exhibit some fragility (see Keane 1992), which has implications for the choice of the prior variances as discussed in Chapter 3. In particular, large prior variances may occasionally produce a draw that could be considered as an outlier. The chain recovers quickly but the precision of the posterior estimates is affected. To avoid getting inconsistent draws, a somewhat tight prior variance can be assumed. However, tight prior variances make the choice of the mean priors critical; hence, mean priors were initiated using the updating mechanism discussed previously by sampling a few thousand iterations from the distributions before starting the sampling scheme per se. Then, another 20,000 iterations were sampled and the last 10,000 were used to generate the means and standard deviations of the estimates. The results in Table 5.2 reveal that on average, many attributes seem to play a role in the decision-making process and are important to understanding why and when respondents are more inclined to select one of the proposed alternatives. Most attributes were assumed to have generic effects; that is, they were assumed to impact each type of institutions similarly, with the exception of an alternative specific constant (ASC) for each institution type. The results also reveal preference heterogeneity in the data, but not for all attributes. Specifically, the impact of each institution-brand exhibits little heterogeneity, which could be due to the domination of institution type effects over brand effects; alternatively, it could be due to the fact that many brands were varied in the experiment, hence more data might be needed to estimate the heterogeneity distribution more precisely. In contrast, type of institution exhibits considerable heterogeneity. That is, on average large and moderately-large institutions were most preferred; moderately-large brands 1 and 3 were the most preferred and Non-traditional institution-brands 2 and 3 were least preferred. Table 5.2 shows the moments of the heterogeneity distribution (mean estimates and variances for all attributes).

Table 5.2: Moments of the heterogeneity distribution

Attributes	Moments of the heterogeneity distribution			
	Mean Estimates	**Standard deviation**	**Variance Estimates**	**Standard deviations**
Large institution	-1.0190	0.1291	0.4074	0.0611
Moderate-Large institution	-1.1210	0.1261	0.3547	0.0583
Small institution	-1.6750	0.1228	0.3740	0.0662
Another type of institution	-1.5650	0.1344	0.6398	0.0919
Large Institution 1	0.0144	0.0166	0.0014	0.0015
Large Institution 2	-0.1529	0.0931	0.0190	0.0224
Large Institution 3	0.0853	0.0691	0.0110	0.0183
Moderate-Large Institution 1	0.5283	0.1124	0.0558	0.0449
Moderate-Large Institution 2	0.0839	0.0641	0.0029	0.0031
Moderate-Large Institution 3	0.4461	0.0984	0.0135	0.0118
Small Institution 1	0.1200	0.0805	0.0221	0.0166
Small Institution 2	0.1034	0.0731	0.0234	0.0362
Small Institution 3	0.0121	0.0118	0.0012	0.0013
Another type of Institution 1	0.1298	0.0799	0.0156	0.0193
Another type of Institution 2	-0.3970	0.1250	0.0216	0.0205
Another type of Institution 3	-0.2926	0.1107	0.0422	0.0444
Hidden attribute	-0.3269	0.0278	0.1149	0.0176
Monthly account fee A	-0.0332	0.0035	0.0010	0.0003
Fee to take money out A	-0.0183	0.0129	7.17E-04	5.82E-04
Fee to put money in A	-0.0038	0.0044	8.80E-04	7.38E-04
Annual interest paid A	0.2712	0.0724	0.2697	0.0828
Monthly account fee B	-0.0085	0.0028	1.28E-04	6.14E-05
Fee to take money out B	-0.0322	0.0168	0.0074	0.0041
Fee to put money in B	-0.0069	0.0068	0.0058	0.0018
Annual interest paid B	0.2351	0.0819	0.4468	0.1146
TA - Fee to transfer within	-0.0857	0.0198	0.0215	0.0083
TA - Fee to transfer outside	-0.0169	0.0129	0.0084	0.0040
Fee to include a line of credit	-0.0071	0.0053	6.29E-04	5.55E-04
CC - Annual fee	-0.0021	0.0004	6.83E-06	2.50E-06
CC - Interest rate	-0.1194	0.0586	0.4636	0.0580
CC - Loyalty scheme Airline 1	-0.0064	0.0148	9.00E-04	9.18E-04
CC - Loyalty scheme Airline 2	0.0562	0.0215	0.0025	0.0016
CC - Loyalty scheme Airline 3	0.0600	0.0264	0.0109	0.0066
IA - Fee to use internet type 1	-0.0130	0.0070	4.91E-04	2.66E-04
IA - Fee to use internet type 1	-0.0162	0.0087	0.0013	0.0010
TA - Fee to use telephone type1	-0.0004	0.0082	1.79E-04	1.34E-04
TA - Fee to use telephone type1	-0.0193	0.0123	0.0011	0.0011
CF - Feature	0.0038	0.0039	8.48E-04	7.49E-04
CF - Attribute 1	-0.0161	0.0111	0.0024	0.0011
CF - Attribute 2	-0.0706	0.0239	0.0102	0.0056
LP - Nearest full service office	-0.0029	0.0038	5.61E-05	3.58E-05
LP - Retailer type 1	-0.0252	0.0206	0.0066	0.0045
LP - Retailer type 2	-0.0745	0.0231	0.0104	0.0062

Earlier it was suggested that an advantage of the proposed estimation approach is the selection procedure that seeks a more parsimonious representation of the covariance structure. Whether this advantage obtains can be assessed by examining the estimated correlations among model parameters in Figure 5.1. That is, larger correlations are darker, such that diagonal elements and off-diagonal correlations larger than 0.8 (or smaller than –0.8) are displayed in black. Correlations between 0.5 and 0.8 (or between –0.5 and –0.8) are in dark grey; light grey is used to represent correlations between 0.2 and 0.5 (or between –0.2 and –0.5). Finally, less important correlations between –0.2 and 0.2 are in white.

Figure 5.2: Pattern of the correlation matrix of the individual level parameters



The structure in Figure 5.2 is somewhat sparse but still presents non-zero correlations, which provides supports for estimating a parsimonious structure. Let us now consider the matrix G to which the selection procedure was applied. The objective was to sample a more parsimonious space while reserving the option to estimate a full covariance structure if needed. Recall that G parameterises the inverse of the covariance matrix Ω. This space corresponds to the partial correlations, which could suggest an even more parsimonious representation. Each element of G_{ij} is zero if and only if the partial correlation between attribute i and j is zero. It should be noted that the primary reason to parameterise Ω^{-1} and not Ω is to make the sampling scheme easier, but it has the additional advantage of potentially revealing more parsimony. Thus, we need to determine whether the approach was able to estimate only a few elements or needed to use the full structure available (i.e., all elements of the triangular matrix G). Figure 5.2 addresses this by depicting the frequency with which the approach used each element of G.

Figure 5.3: Pattern of the probabilities to select/estimate elements of G

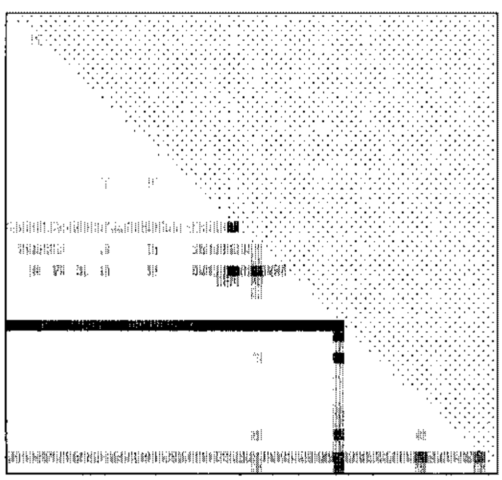


Figure 5.2 demonstrates that the method was successful in estimating a more parsimonious structure. Important elements, with a probability superior to 0.8, are in black while elements whose probability is between 0.8 and 0.5, and between 0.5 and 0.2 are in dark and light grey, respectively. White depicts elements with a probability of being selected/estimated of less than 0.2.

Thus, more than 75% of the G elements were set to zero in more than 90% of the iterations used, which is not altogether surprising because estimation of so many elements (even if they all matter) would require considerable information to obtain relatively precise estimates; hence this parsimonious structure partly reflects lack of data. Obviously, elements that clearly differ from zero are selected and estimated at each iteration, and hence have a high probability; unimportant elements are discarded (i.e. set to zero) and reflected by low probabilities. It is worth noting that at each iteration a different model structure might be (and generally is) estimated; hence at the end of the sampling process the model estimates represent an average over all considered structures. This process is termed "model averaging" and weights each model/structure by its frequency (i.e., its importance).

Finally, the estimation approach permitted a test of a common assumption regarding the first level of the hierarchy, namely that the errors are IID. Table 5.3 presents the estimates of the covariance matrix Σ in matrix form; the lower triangle represents the estimates of the covariances and the upper triangle contains their respective standard deviations. The standard deviations of the diagonal elements are in parenthesis. Note that the last element Σ_{55} of the diagonal was estimated in a first stage in which a diagonal structure was imposed. All covariance elements were then estimated in a second stage with $\Sigma_{55} = 1.811$.

Table 5.3: Estimates of the covariance matrix Σ

Large Institutions	1 - fixed	(7.76E-06)	(2.61E-06)	(3.14E-06)	0
Moderate-Large Institutions	-1.86E-07	0.314 (0.012)	(4.99E-06)	(6.27E-06)	0
Small Institutions	-4.48E-08	1.32E-07	0.279 (0.011)	(5.07E-06)	0
Non-traditional Institution	-6.21E-08	2.18E-07	2.08E-07	0.312 (0.012)	0
Current	0	0	0	0	1.811 - fixed

The results in Figure 5.3 suggest that the assumption of zero covariances can be retained at this level, but the diagonal elements do not appear to be identically distributed. The probabilities that the covariance elements were different from zero were smaller than 0.2%, which means that the estimation procedure almost never considered a non-diagonal structure, probably due to the fragility of MNP models mentioned earlier (see also Chapter 3). If the data cannot reveal the structure, the selection procedure sets the elements to zero at almost every iteration.

In the present case a diagonal structure was estimated but the priors were somewhat tight for this covariance matrix; hence this structure can be viewed as lying between an IID and a fully diagonal model that suggest an important conclusion. That is, the variance of the unobserved component is larger for the "I'd keep my current account" option, which is consistent with the results of Dellaert, Borgers and Timmermans (1996), who found that the random components of specific alternatives differed from that of the constant option (a "none" option). In conclusion, a diagonal structure may be all that is required to model the first level of the hierarchy in this dataset.

5.5 Conclusions

This chapter developed an MNP model that can capture the full covariance structure of individual preferences if necessary; but also allows one to select and estimate a more parsimonious structure instead. A similar selection procedure was used to obtain a parsimonious representation of the covariance structure of the utilities. The advantages of element selection procedures like this are that researchers do not need to rely on an a priori structure, but instead can estimate fewer structural elements if the data support this. The proposed approach was illustrated by an empirical application involving choice of financial institution types for a certain type of financial services account, and the approach successfully found a more parsimonious structure for the heterogeneity covariance matrix and a non-IID diagonal error structure.

Finally, we showed how to exploit preference information additional to that contained in typical "most preferred option" choice experiment responses. Such extra information can be used to better identify and estimate individual utility functions; for example, if consumers choose constant options like “none” or “keep my current” option with high frequencies, the additional information permits one to estimate preference trade-offs and avoid potential loss of information due to sparse responses. Responses to "which option would you seriously consider", "which option would be your second best choice", or eventually a complete ranking of options provide a safety net that allows one to obtain aggregate as well as individual estimates. The empirical illustration showed that one can combine such information to better understand consumer preferences, even in cases in which they may not show significant interest the options offered in an experiment.

Unobserved attributes may impact similar brands (or types of brands in the empirical example in this chapter) in a similar way; hence models that relax IID-error assumptions may provide more accurate descriptions of market structure. Similarly, uncovering the covariance structure of individual preferences allows one to better identify potential differences between customers. In summary, models such as the one presented in this research allow marketers to obtain a more precise picture of market structure.

Future research should investigate the tradeoffs involved in estimating assumption-free covariance structures compared with more restrictive models like factor-analytic probit for different types of markets (e.g., Elrod and Keane 1995). Selection procedures could also be applied directly to individual preferences to select sets of attributes that individuals use to make choices. That is, if information on many attributes is provided to respondents, it is unlikely that they will use all this information. Heterogeneity in such data confirms that individuals have different preferences for some of the attributes; hence estimates can be improved if unimportant attributes can be discarded at an individual level. The latter could permit one to find segments of customers who use the same set of attributes to make choices, even if their preferences for the attributes differ.

Chapter 6

Discussion and Conclusions

This thesis developed new Bayesian procedures to estimate the parameters of Multinomial Probit (MNP) models, with particular attention given to the error covariance matrix. That is, the MNP structure is highly flexible and can account for errors that are not independently and identically distributed (IID), resulting in choice probabilities that do not satisfy the independence of irrelevant alternatives (IIA) property of models like Multinomial Logit (MNL). For example, MNL results from assuming that the errors are IID extreme value type I; thus the IIA property of MNL is a consequence of that assumption. In marketing this property has been found to be restrictive, hence researchers have focused attention of models without that restriction, which in turn has lead to the development of improved estimation methods for MNP models.

Despite MNP's greater flexibility and behavioural appeal, estimation remains somewhat challenging, particularly estimation of the error covariance matrix. The identification issues now are well known, but estimates can be fragile in the sense that the likelihood is flat near its maximum. Fragility of estimates depends on data, which means that some data sets yield good estimates, whereas others experience convergence problems. One approach to this problem is to use Bayesian estimation methods to develop more parsimonious representations if data support and/or require it. The latter approach can be used to achieve parsimony by using an element selection procedure applied to a decomposition of the covariance matrix that selects and estimates only the important elements. The Gibbs sampling scheme for this problem was presented in Chapter 3, and illustrated by means of an empirical application using a revealed preference data set that has been examined by other researchers (data supplied by the Limdep Econometric Software). The proposed approach found the covariance matrix to be diagonal, hence the sampler set all covariance elements to zero.

Chapter 4 investigated the ability of particular types of choice experiments to estimate MNP model parameters. That is, Chapter 4 addressed the following issues: 1) whether choice experiments that are constructed to estimate MNL parameters of interest as precisely as possible, but were not specifically designed to estimate MNP parameters, can exploit the more complex structure of MNP models; 2) how good an approximation is MNL and what is gained by estimation of more complex MNP models? Typically, one tries to design choice experiments so that each parameter can be estimated independently of all others. Parameters that are specific to each alternative supposedly can be estimated from designs that are orthogonal within- and between-alternatives while generic parameters (i.e. each attribute impacts each alternative identically) can be estimated from experiments that are orthogonal only within-alternatives.

A Monte Carlo simulation approach was used to investigate and compare MNL and MNP estimates under conditions in which the true errors were non-IID. The results showed that estimated parameters are close to the true values for the case of generic parameters. Importantly and surprisingly, it was found that the generic slope parameters were very precisely estimated for both MNP and MNL models. However, when the alternative-specific constants (ASC's) were estimated using MNL, they consistently were biased, which affects the model predictions, but not strategic conclusions based on slope estimates. Satisfactory MNP covariance estimates were achieved with both types of choice experiments but only in the case of generic parameters, and covariance matrices were not well-estimated if alternative-specific parameters also were estimated. The objective of within-alternative designs is to maximise statistical efficiency of MNL model parameters by maximising the number of between-attribute differences. A comparison of within- with within- and between-alternative orthogonal attribute designs for the case of generic parameters revealed little difference in the two designs in terms of parameter efficiency, but the within- and between-designs lead to significantly more scenarios. In the case of alternative-specific parameters, the within- and between-designs not only did not perform as well as expected, but there also was little difference in the quality of the estimates produced by these more complex designs and the simpler within-only designs. Thus, future research should consider whether within-only designs that maximise the number of attribute differences will perform as well if not better systematically than within- and between-designs.

Chapter 5 considered the distribution of differences in individual preferences, or the distribution of preference heterogeneity. Estimation of this distribution generally involves estimation of a variance-covariance matrix of attribute parameters. The number of elements of this matrix increases quadratically with the number of attributes; hence, a hierarchical model was proposed to achieve parsimony for large numbers of attributes.

Models that provide more information are usually desirable but the amount of data available dictates the level of detail that can be obtained; moreover, the more complex a model, the greater the likelihood of overfitting the data. Chapter 5 proposed two solutions to the problem of complexity. The first one is straightforward: more information is needed. In choice experiments the option "not to choose" is usually given to respondents to make the task more realistic. Inclusion of such realistic options is important, but estimation problems arise if subjects in choice experiments choose such options a large proportion of the time because there is less information available to estimate individual preference trade-offs.

Chapter 5 proposed one solution to this problem that involves asking multiple preference questions to ensure that respondents will express preferences for the designed choice options. One such question included a very unappealing constant option to encourage respondents to tradeoff attributes of designed options and select one of them. Another asked "consideration questions" (e.g., "which options would you seriously consider?") that yield information on weak rankings. Complete rankings are another possibility; but increase in cognitive difficulty as numbers of choice options increase; hence one might consider a partial ranking instead.

Whatever solution one adopts, the point of view taken in Chapter 5 was that complex models either require more information per choice set or more choice sets per respondent. More importantly, Chapter 5 proposed a mathematical solution for complexity in the form of a model that searches for more parsimonious representations of error structures by recognising that some elements should be excluded from estimation. This can arise either because such elements do not matter or because there is not sufficient information in the data to produce reasonably precise estimates. The proposed solution uses Bayesian element selection procedures to estimate a parsimonious covariance matrix using "most preferred" and "would consider" experimental data. Application to an empirical example involving financial service accounts demonstrated that the covariance structure was sparse and that many elements were irrelevant.

Many extensions are possible to this estimation approach. For example, element and/or variable selection methods have the advantage that complex models can be proposed but the estimated structure ultimately depends on the data, which may result in structures as simple as MNL. Larger covariance matrices that allow for correlation among choice sets also could benefit from such element selection methods, and it should be possible to extend the approach to variable selection at the individual level. For data sets of the size typically available to marketing academics and practitioners, it generally should be the case that variance-covariance matrices will be sparse, but such methods should be able to account for potentially important correlations at different levels of aggregation.

The results in this thesis as well as previous research suggest that Bayesian methods should have a promising future in marketing research and related applications. In particular, variable selection, covariance element selection and model averaging provide powerful ways to estimate complex models. However, the issues of model complexity should not be viewed as solely a modeling issue because the potential complexity of a model also depends on the data available for estimation. Thus, choice experiments provide potentially reliable sources of data for model estimation, but it is necessary to understand what can and cannot be estimated from such particular experiments. That is, more research in statistical design theory is needed to identify designs that can support estimation of more complex model structures like MNP. However, more research also is needed on the modeling side to permit better identification of estimation issues related to the use of experimental data.

References

Adamowicz, W. L., Louviere, J., Williams, M. (1994), "Combining Revealed and Stated Preference Methods for Valuing Environmental Amenities", *Journal of Environmental Economics & Management,* 26, 271-92

Albert, J. H., Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data", *Journal of the American Statistical Association,* 88, 669-79

Alderson, W. (1983), "The Heterogeneous Market and the Organized Behavior System" in *Marketing Theory: The Philosophy of Marketing Science*, ed. S. D. Hunt, 292-313, Richard Irwin Inc.

Allenby, G. M., Ginter, J. L. (1995), "Using Extreme to Design Products and Segments Markets", *Journal of Marketing Research,* 32, 392-403

Allenby, G. M., Arora, N., Ginter, J. L. (1998), "On the Heterogeneity of Demand", *Journal of Marketing Research,* 35, 384-389

Allenby, G. M., Rossi, P. E. (1999), "Marketing models of consumer heterogeneity", *Journal of Econometrics,* 89, 57-78

Amemiya, T. (1981), "Qualitative Response Models: A Survey", *Journal of Economic Literature,* XIX, 1483-1536

Anderson, D. A., Wiley, J. B. (1992), "Efficient Choice Set Designs for Estimating Availability Cross-Effects Models", *Marketing Letters,* 3, 357-370

Atkinson, A. C., Donev, A. N. (1992), *Optimum experimental designs*. Oxford statistical science series; 8., Oxford [England], New York, Clarendon Press; Oxford University Press

Beggs, S., Cardell, S., Hausman, J. (1981), "Assessing the potential demand for electric cars", *Journal of Econometrics,* 17, 1-20

Ben-Akiva, M., Lerman, S. R. (1985), *Discrete choice analysis : theory and application to travel demand.* MIT Press series in transportation studies; 9., M. Press Cambridge, Mass. MIT Press

Ben-Akiva, M., Morikawa, T. (1990), "Estimation of Switching Models from Revealed Preferences and Stated Intentions", *Transportation Research A,* 24A, 485-495

Bernardo J. M., Smith A.F.M., (1994), "Bayesian Theory", Chichester Eng; NewYork: Wiley

Bolduc, D., Lacroix, G., Muller, C. (1996), "The Choice of Medical Providers in Rural Benin: A Comparison of Discrete Choice Models", *Journal of Health Economics,* 15, 477-498

Borsch-Supan, A., Hajivassiliou, V. A. (1993), "Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models", *Journal of Econometrics,* 58, 347-68

Brazell, J. D., Louviere, J. J. (1998), "Length Effects in Conjoint Choice Experiments and Surveys: An Explanation Based on Cumulative Cognitive Burden", Department of Marketing, University of Sydney, Australia

Brownstone, D., Train, K. (1999), "Forecasting New Product Penetration with Flexible Substitution Patterns", *Journal of Econometrics,* 89, 109-129

Brydon, K., Louviere, J. J. (1997), "Demand Effects in Experimental Choice", Department of Marketing, University of Sydney, Australia

Bucklin, R. E., Lattin, J. M. (1991), "A Two-State Model of Purchase Incidence and Brand Choice", *Marketing Science,* 10, 24-39

Bunch, D. (1991), "Estimability in the Multinomial Probit Model", *Transportation Research B,* 25B, 1-12

Bunch, D. S., Louviere, J. J., Anderson, D. (1996), "A Comparison of Experimental Design Strategies for Choice-Based Conjoint Analysis with Generic-Attribute Multinomial Logit Models", Graduate School of Management UC Davis, California

Carroll, J. D., Green, P. E. (1995), "Psychometric methods in marketing research: I. Conjoint analysis", *Journal of Marketing Research,* 32, 385-391

Carson, R., Louviere, J. J., Anderson, D. A., Arabie, P., Hensher, D., Johnson, R., Kuhfeld, W., Steinberg, D., Swait, J., Timmermans, H., Wiley, J. (1994), "Experimental Analysis of Choice", *Marketing Letters,* 5, 351-368

Casella, G., George, E. I. (1992), "Explaining the Gibbs Sampler", *The American Statistician,* 46, 167-174

Chapman, R. G., Staelin, R. (1982), "Exploiting Rank Ordered Choice Set Data Within the Stochastic Utility Model", *Journal of Marketing Research,* 19, 288-301

Chib, S., Greenberg, E., Chen, Y. (1998), "MCMC Methods for Fitting and Comparing Multinomial Response Models", manuscript.

Chintagunta, P. K., Jain, D. C., Vilcassim, N. J. (1991), "Investigating Heterogeneity in Brand Preferences in Logit Models for Panel Data", *Journal of Marketing Research,* 28, 417-428

Chintagunta, P. K. (1992), "Estimating a Multinomial Probit Model of Brand Choice Using the Method of Simulated Moments", *Marketing Science,* 11, 386-407

Chintagunta, P. K. (1993), "Investigating Purchase Incidence, Brand Choice and Purchase Quantity Decisions of Households", *Marketing Science,* 12, 184-208

Claycamp, H. J., Massy, W. F. A. (1968), "Theory of Market Segmentation", *Journal of Marketing Research,* 5, 34-45

Dellaert, B., Borgers, A., Timmermans, H. (1996), "Conjoint Choice Models of Joint Participation and Activity Choice", *International Journal of Research in Marketing,* 13, 251-264

Dey, A., Mukerjee, R. (1999), *Fractional Factorial Plans.* Probability and Statistics, W. Series John Wiley & Sons

Diener, C., Dellaert, B. G. C. (1997), "The Effect of Involvement on Choice Consistency", Department of Marketing, University of Sydney, Australia

El-Helbawy, A. T., Bradley, R. A. (1978), "Treatment Contrasts in Paired Comparisons: Large-Sample Results, Applications, and Some Optimal Designs", *Journal of the American Statistical Association, (Theory and Methods),* 73, 831-839

Elrod, T., Louviere, J. J., Davey, K. S. (1992), "An empirical comparison of ratings-based and choice-based conjoint models", *Journal of Marketing Research,* 29, 368-377

Elrod, T., Keane, M. P. (1995), "A Factor-Analytic Probit Model for Representing the Market Structure in Panel Data", *Journal of Marketing Research,* 32, 1-16

Frank, R. E., Massy, W. F., Wind, Y. (1972), *Market Segmentation.* Englewood Cliffs, NJ, Prentice-Hall, Inc.

Gelfand, A. E., Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities", *Journal of the American Statistical Association,* 85, 398-409

Gelman, A. (1995), "Inference and Monitoring Convergence", In *Markov Chain Monte Carlo in Practice* (eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter), 131-143. London: Chapman & Hall

Gelman, A., Rubin, D. B. (1992), "Inference from iterative simulation Using Multiple Sequences", Statistical Science, 7, 457-511

George, E. I., McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling", *Journal of the American Statistical Association,* 88, 881-889

George, E. I., McCulloch, R. E. (1997), "Approaches to Bayesian variable selection", *Statistica Sinica,* 7, 339-374

Geweke, J. (1991), Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraints Probabilities. *Proceedings of the Twenty-Third Symposium on the Interface Between Computing Science and Statistics*, ed. *E. Kermamidas*, 571-578,

Geweke, J., Keane, M., Runkle, D. (1994), "Alternative Computational Approaches to Inference in the Multinomial Probit Model", *Review of Economics & Statistics,* 76, 609-32

Green, P. E., Rao, V. R. (1971), "Conjoint Measurement for Quantifying Judgmental Data", *Journal of Marketing Research,* 8, 355-363

Green, P. E., Wind, Y. (1973), *Multiattribute Decisions in Marketing: A Measurement Approach*. Hinsdale, Ill., Dryden Press

Green, P. E., Srinivasan, V. (1978), "Conjoint analysis in consumer research: issues and outlook", *Journal of Consumer Research,* 5, 103-123

Green, P. E., Srinivasan, V. (1990), "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice", *Journal of Marketing,* 54, 3-19

Grover, R., Srinivasan, V. (1989), "An Approach for Tracking Within-Segment Shifts in Market Shares", *Journal of Marketing Research,* 26, 230-236

Guadagni, P. M., Little, J. D. C. (1983), "A Logit Model of Brand Choice Calibrated on Scanner Data", *Marketing Science,* 2, 203-238

Haaijer, R. E., Wedel, M., Vriens, M., Wandsbeek, T. J. (1998), "Utility Covariances and Context Effects in Conjoint MNP Models", *Marketing Science,* 17, 236-52

Hajivassiliou, V., Ruud, P. (1994), "Classical Estimation Methods for LDV Models Using Simulation" in *Handbook of Econometrics,* ed. R. E. a. D. McFadden, vol. 4, 2383-2441, Amsterdam: North Holland

Hajivassiliou, V. A., McFadden, D., Ruud, P. E. (1996), "Simulation of Multivariate Normal Rectangle Probabilities and Their Derivatives: Theoretical and Computational Results", *Journal of Econometrics,* 72, 85-134

Hajivassiliou, V., McFadden, D. (1998), "The Method of Simulated Scores for the Estimation of LDV Models", *Econometrica,* 66, 863-896

Hensher, D. A., Bradley, M. (1993), "Using stated response data to enrich revealed preference discrete choice models", *Marketing Letters,* 4, 139-152

Hensher, D. A., Louviere, J. J., Swait, J. (1999), "Combining Sources of Preference Data", *Journal of Econometrics,* 89, 197-221

Horowitz, J. (1980), "The Accuracy of the Multinomial Logit Model as an Approximation to the Multinomial Probit Model of Travel Demand", *Transportation Research B,* 14B, 331-341

Horowitz, J. L., Louviere, J. J. (1995), "What is the role of consideration sets in choice modeling?", *International Journal of Research in Marketing,* 12, 39-54

Huber, J., Puto, C. (1982), "Market Boundaries and Product Choice: Illustrating Attraction and Substitution Effects", *Journal of Consumer Research,* 10, 31-44

Huber, J., Zwerina, K. (1996), "The importance of utility balance in efficient choice designs", *Journal of Marketing Research,* 33, 307-317

Johnson, R. M., Orme, B. K. (1996), "How Many Questions Should You Ask In Choice-Based Conjoint Studies", *ART Forum*, Beaver Creek, CO

Kamakura, W. A., Russell, G. J. (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structuring", *Journal of Marketing Research,* 379-90

Keane, M. P. (1992), "A Note on Identification in the Multinomial Probit Model", *Journal of Business & Economic Statistics,* 10, 193-200

Keane, M. (1994), "A Computationally Practical Simulation Estimator for Panel Data", *Econometrica,* 65, 95-116

Kuhfeld, W. F., Tobias, R. D., Garratt, M. (1994), "Efficient Experimental Design With Marketing Research Applications", *Journal of Marketing Research,* 21, 545-557

Lancaster, K. (1966), "A New Approach to Consumer Theory", *Journal of Political Economy,* 74, 132-157

Layton, D. F., Lee, T. S. (1998), "From Ratings to Rankings: The Econometric Analysis of Stated Preference Ratings Data", Department of Enviornmental Science and Policy, University of California, Davis

Lerman, S. R., Louviere, J. J. (1979), "On the use of direct utility assessment to identify the functional form of utility and destinations choice models", *Transportation Research Record,* 673, 78-86

Little R., and D. Rubin (1987), "Statistical Analysis with Missing Data", New York, Wiley.

Lindley, D. V., Smith, F. M. (1972), "Bayes Estimates for the Linear Model", *Journal of the Royal Statistical Association B,* 35, 370-384

Louviere, J. J. (1979), "Modeling Individual residential preferences: a totally disaggregate approach", *Transportation Research A,* 13, 374-384

Louviere, J. J., Meyer, R. J. (1981), "A Composite Attitude-Behavior Model of Traveler Decision Making", *Transportation Research B,* 15, 411-420

Louviere, J. J., Hensher, D. A. (1983), "Using discrete choice models with experimental design data to forecast consumer demand for a unique cultural event", *Journal of Consumer Research,* 10, 348-361

Louviere, J. J., Woodworth, G. (1983), "Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data", *Journal of Marketing Research,* 20, 350-367

Louviere, J. J. (1988), *Analysing Individual Decision Making: Metric Conjoint Analysis.* Sage University Series on Quantitative Applications in the Social Sciences, ed., 67. Newbury Park, CA, Sage Publications, Inc.

Louviere, J., Anderson, D., White, J., T., E. (1990), "Predicting Preferences for New Product Configurations: A High-Tech Example" in *Modelling The Innovation: Communications, Automation and Information Systems*, ed. M. Carnevale, M. Lucertini and S. Nicosia, Elsevier Science, North-Holland

Louviere, J. J., Fox, M. F., Moore, W. L. (1993), "Cross-Task Validity Comparisons of Stated Preference Choice Models", *Marketing Letters,* 4, 205-213

Louviere, J. J. (1994), "Conjoint Analysis" in *Advanced Marketing Research*, ed. R. Bagozzi, Blackwell Publishers

Louviere, J. J., Hensher, D. A., Swait, J. (2000), *Stated Preference Methods: Analysis and Applications*. Cambridge, Cambridge University Press

Luce, R. D. (1959), *Individual Choice Behaviour: A Theoretical Analysis.* New York, Wiley

Luce, D. R., Tukey, J. W. (1964), "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement", *Journal of Mathematical Psychology,* 1, 1-27

Magnus, J., Neudecker, H. (1998), *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Series in Probability and Mathematical Statistics, Wiley Amsterdam, John Wiley and Sons

McCulloch, R. E., Rossi, P. E. (1994), "An Exact Likelihood Analysis of the Multinomial Probit Model", *Journal of Econometrics,* 64, 207-40

McCulloch, R., Polson, N., Rossi, P. E. (2000), "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters", *Journal of Econometrics,* Forthcoming

McFadden, D. (1973), "Conditional logit analysis of qualitative choice behaviour" in *Frontiers in Econometrics*, ed. P. Zarembka, 105-142, New York: Academic Press

McFadden, D., Train, K. (2000), "Mixed MNL Models for Discrete Response", *Journal of Applied Econometrics,* Forthcoming

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A., Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines", *Journal of Chemical Physics,* 21,

Mitchell, T. J., Beauchamp, J. J. (1988), "Bayesian variable selection in linear regression", *Journal of the American Statistical Association,* 83, 1023-1032

Oppewal, H., Louviere, J. J., Timmermans, H. J. P. (1994), "Modeling hierarchical conjoint processes with integrated choice experiments", *Journal of Marketing Research,* 31, 92-105

Raftery, A. E., Madigan, D., Hoeting, J. (1997), "Bayesian model averaging for linear regression models", *Journal of the American Statistical Association,* 92, 179-191

Revelt, D., Train, K. (1999), "Customer-Specific Taste Parameters and Mixed Logit", Berkeley, California: Department of Economics, University of California, Berkeley

Rossi, P. E., Allenby, G. M. (1993), "A Bayesian Approach to Estimating Household Parameters", *Journal of Maketing Research,* 30, 171-182

Severin, V., Louviere, J. J., Finn, A. (2000), "Stability of Retail Shopping Choices Over Time and Across Countries", under review at *Journal of Retailing*

Simonson, I., Tversky, A. (1992), "Choice in context: Tradeoff contrast and extremeness aversion", *Journal of Marketing Research,* 29, 281-295

Smith, W. R. (1956), "Product Differentiation and Market Segmentation as Alternative Marketing Strategies", *Journal of Marketing Research,* 21, 3-8

Smith, M., Kohn, R. (1996), "Nonparametric regression using Bayesian variable selection", *Journal of Econometrics,* 75, 317-344

Smith, M., Kohn, R. (1999), "Bayesian parsimonious covariance matrix estimation", Preprint

Street, D. J., Bunch, D. S., Moore, B. (1999), "Optimal Designs for 2k Paired Comparison Experiments", School of Mathematical Sciences, University of Technology, Sydney, Australia

Street, D. J., Bunch, D. S., Moore, B. (1999), "Optimal Designs for 2k Paired Comparison Experiments", School of Mathematical Sciences, University of Technology, Sydney, Australia

Swait, J., Louviere, J. (1993), "The role of the scale parameter in the estimation and comparison of multinomial logit models", *Journal of Marketing Research,* 30, 305-314

Swait, J., Louviere, J. J., Williams, W. (1994), "A sequential approach to exploiting the combined strengths of SP and RP data: Application to freight shipper choice", *Transportation,* 21, 135-152

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

Thurstone, L. L. (1927), "A Law of Comparative Judgement", *Psychological Review,* 34, 273-286

Train, K. (1999), "Halton Sequences for Mixed Logit", Berkeley, California: Department of Economics, University of California, Berkeley

Vriens, M., Oppewal H., M., W. (1998), "Ratings-based versus choice-based latent class conjoint models-An empirical comparison", *Journal of the Market Research Society,* 40, Jul 1998

Wedel, M., Kamakura, W. A. (1998), *Marketing Segmentation: Conceptual and Methodological Foundations.* Boston: Kluwer Academic Publishers

Wiley, J. B., Anderson, D. A. (1992), "On the Design of Efficient Choice Set Experiments for Estimating Competing Effects Models", *Marketing Letters,* 3, 357-370

Wind, Y. (1978), "Issues and Advances in Segmentation Research", *Journal of Marketing Research,* 15, 317-337

Wittink, D. R., Cattin, P. (1989), "Commercial use of conjoint analysis: An update", *Journal of Marketing,* 53, 91-96

Yeoh, A., Uldry, P.-F., Louviere, J., Burke, S. (1998), "Does Missing Information Affect Means or Variances of Error Distributions?", *Marketing Science* Conference, INSEAD, France