

University of Groningen

## Essays on Customization Applications in Marketing

Adiguzel, Feray

**IMPORTANT NOTE:** You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

*Document Version*

Publisher's PDF, also known as Version of record

*Publication date:*

2006

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*

Adiguzel, F. (2006). Essays on Customization Applications in Marketing s.n.

**Copyright**

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

**Take-down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

**ESSAYS ON  
CUSTOMIZATION APPLICATIONS  
IN MARKETING**

Published by: Labyrinth Publications  
P.O. Box 334  
2984 AX Ridderkerk  
The Netherlands  
Tel: + 31 180 463 962

Printed by:  Offsetdrukkerij Ridderprint B.V., Ridderkerk

© 2006, Feray Adıgüzel

All rights reserved. No part of this publication may be reprinted or utilized in any form or by any electronic, mechanical or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without written permission from the copyrights owner.

ISBN 90-5335-089-6



Rijksuniversiteit Groningen

**ESSAYS ON  
CUSTOMIZATION APPLICATIONS  
IN MARKETING**

Proefschrift

ter verkrijging van het doctoraat in de  
Economische Wetenschappen  
aan de Rijksuniversiteit Groningen  
op gezag van de  
Rector Magnificus, dr. F. Zwarts,  
in het openbaar te verdedigen op  
maandag 19 juni 2006  
om 13.15 uur

door

**FERAY ADIGÜZEL**

geboren op 30 maart 1974  
te İmranlı (Turkije)

**Promotor:**

Prof. dr. M. Wedel

**Copromotor:**

Dr. J. Zhang

**Beoordelingscommissie:**

Prof. dr. T.H.A. Bijmolt

Prof. dr. R.T. Frambach

Prof. dr. P.C. Verhoef

**ISBN 90-5335-089-6**



## Acknowledgements

This thesis is the result of projects conducted at the University of Groningen and Stephen M. Ross School of Business at the University of Michigan. I learned many things both scientifically and nonscientifically during my years as a Ph.D student and I'm quite grateful for the experiences I had during this time. This work could not have been conducted without the help and support of many people.

First, I would like to thank my supervisor Prof. Dr. Michel Wedel for giving me the opportunity to accomplish this work at the Department of Marketing in the University of Groningen and the University of Michigan, as well as for his guidance and help during these past years. His enthusiasm, his support and his input encouraged me to finish this dissertation. Michel: I was lucky to work with you on my Ph.D. because you are not only the best model of a good researcher, but at the same time the best model of a good person.

I would like to thank Prof. Dr. Jie Zhang for agreeing to be a member of my Ph.D committee. Through my stay in UM, I had the opportunity to get to know her better and to work for her as a research assistant. I found that she is a very helpful and thoughtful person and I'm glad to have had the opportunity to work on a paper together with her, which comprises Chapter Four of this thesis.

I'm grateful to Prof. Dr. Fred Feinberg for the opportunity to work on the "Antilium" project which resulted in a paper. I greatly enjoyed working with him and I learned many things from him, in particular about American culture and American English.

I would like to thank the SOM graduate school for their academic and financial support. I'm especially grateful to Dirk Pieter van Donk from the SOM graduate school for his help. Not only was he very quick in responding to any questions I had during my stay in Ann Arbor, he was also very supportive of me during the process of finishing this dissertation.

I would like to thank the other members of the reading committee, Prof. Dr. Tammo Bijmolt, Prof. Dr. Ruud Frambach, and Prof. Dr. Peter Verhoef, for agreeing to be members of my Ph.D. committee and for their comments and suggestions.

I would like to thank my friend, my erstwhile officemate and paranimph, Marije Teerling, for all office and nonoffice talks and help with my defense arrangements. Special thanks go to my other paranimph Csilla Horvath. Csilla: As you said earlier, it is really not "lehetetlen."

I would like to thank everyone in the Marketing Department of University of Michigan for their hospitality and support. I greatly enjoyed visiting them and am proud of being a member of their marketing department for the last three years.

I would like to thank my colleagues at the Department of Marketing and Marketing Research at the University of Groningen for the friendly atmosphere they provided during my stay. Special thanks go to my Dutch friends Josephine Woltman Elpers, Albert

van Dijk, Peter Ebbes, and Ralf van der Lans, my Portuguese friend and officemate José Dias for the intellectual discussions we had and my Czech coauthor Tomas Sobotka for the opportunity of working on the paper together. I'm grateful to Rutger van Oest for his support and for providing the Dutch summary of the dissertation, especially for being prompt with the translation and for the high level accuracy of his work.

I would like to thank my friends in Michigan: I am honored to have made all of your acquaintance. I will mention some of them: Anna Camacho, Jing Xu, Francisco J. Aragón Artacho, U of M Marketing Ph.D. students, and my friends from the Michigan Argentine Tango Club.

Finally, I would like thank to my family for their constant support, good wishes and prayers: my parents, my sisters Tülay and Tuğba, my brothers Canhasan and Tolga Adıgüzel. I always feel wherever I go, they are with me.

Ann Arbor, Michigan, 2006





# Contents

<b>Acknowledgements</b>	<b>1</b>
<b>Contents</b>	<b>5</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 Customization in Marketing	1
1.2 Marketing Mix Customization	3
1.2.1 Customizing Product	4
1.2.2 Customizing Price	5
1.2.3 Customizing Communication	5
1.2.4 Customizing Distribution	6
1.2.5 Customizing After-Sales Support and Costs	7
1.3 Problem Delineation:	7
1.4 Motivation of Essay 1:	10
1.5 Motivation of Essay 2:	14
1.6 Outline of the Dissertation	17
<b>Chapter 2 Decision Making under Uncertainty</b>	<b>21</b>
2.1 Introduction	21
2.2 Bayesian Analysis and Marketing Decisions	30
<b>Chapter 3 Split Questionnaire Design</b>	<b>37</b>
3.1 Introduction	37
3.1.1 Motivation	37
3.1.2 Outline of the Chapter	40
3.2 Constructing the Split Questionnaires	42
3.3 Data Missing by Design	43
3.3.1 Two-Stage Designs	48
3.3.2 Matrix Sampling Design	48
3.3.3 Time Sampling Design	50

3.3.4	Subsampling or Multistage Sampling	52
3.3.5	Data Fusion	52
3.3.6	Incomplete Block Design	55
3.4	Measuring Information Loss	59
3.4.1	Optimal Split Questionnaires Using KLD	59
3.5	Identification Issues in Constructing SQD	62
3.6	Design Generating Algorithm	65
3.6.1	Generating Within-Block Designs	67
3.7	Multiple Imputations with Gibbs Sampling	69
3.8	Estimation of the Fraction of Missing Information	71
3.9	Simulation Studies	72
3.10	Empirical Data Application	76
3.10.1	Between-Block Designs	79
3.10.2	Within-block Designs	81
3.11	Field Study	85
3.12	Conclusion	90
3.13	Appendix	93
3.13.1	KL-Distance for Mixed Data	93
3.13.2	Missing Information Principle	95
3.13.3	Figures	98

## **Chapter 4 Promotion Customization across Multiple Categories 103**

4.1	Introduction	103
4.2	Literature Review	108
4.3	Methodology	110
4.3.1	The Model	110
4.3.2	Consumer Response Model	111
4.3.3	Individual Level Heterogeneity	114
4.3.4	Joint Model (Hierarchical Multivariate Type-2 Tobit Model)	115
4.4	Estimation with MCMC	116
4.4.1	Gibbs Sampling	117
4.5	Customized Promotions Design	120
4.5.1	Modified Fedorov Algorithm	125
4.6	Synthetic Data Results for Model Estimation	126
4.6.1	No Covariance between Incidence and Expenditure, No Individual Heterogeneity	127
4.6.2	With Covariance between Incidence and Expenditure, No Individual Heterogeneity	128
4.6.3	With Covariance between Incidence and Expenditure, and Individual Heterogeneity	129
4.7	Data Description	132
4.7.1	Model Specification and Variable Definition	141
4.8	Results and Discussion	142
4.9	Optimization Results	147
4.10	Conclusion	155

<b>Chapter 5 Conclusion and Discussion</b>	<b>159</b>
5.1 Introduction	159
5.2 Summary and Conclusions	159
5.3 Limitations and Future Research	163
5.3.1 Split Questionnaires	164
5.3.2 Customization	169
<b>References</b>	<b>173</b>
<b>Subject Index</b>	<b>189</b>
<b>Author Index</b>	<b>193</b>
<b>Samenvatting (Summary in Dutch)</b>	<b>197</b>
Samenvatting en Conclusies	197





# Chapter 1

## Introduction

### 1.1 Customization in Marketing

True one-to-one customization has begun to be realized by more companies through the migration of marketing to the online environment. Companies are changing their marketing strategies from being seller-centric to being buyer-centric. For this purpose, they develop methods and strategies to customize marketing mix instruments i.e. product, purchase price, communication, distribution and logistics, and after-sales support and cost (Rust and Verhoef, 2005). As a result, we observe that customers are becoming active participants in the product development, purchase, and consumption processes in the digital marketplace. For example, Dell computer designs a personal computer based on the specifications which are set by customers from a choice menu. In the car industry, GM and Chrysler are examples of companies engaging in product and price customization. Wind and Rangaswamy (2001) call this emerging paradigm “customerization” and describe it as “a call to everyone in the marketing profession to rise to a new standard of interacting with customers and building relationships with them.” This new paradigm merges mass customization, one-to-one marketing strategies, and focuses interest on the firm decisions of ‘whom to target, when and with what’ and on the customer decisions of ‘whether, what, when and where to buy’. In sum, with the adaptation of one-to-one marketing strategies to the Internet, marketing strategies are becoming more individually oriented. Such strategies often require little prior information about customers, and even the product itself

can be manufactured after consumers tell the company what they want to buy.

Businesses have begun to develop databases that allow them to approach customers on an individual basis by customizing their ways of introducing, providing, and delivering products and services to the customers. Nowadays, especially in business-to-business markets, many firms are starting to involve their customers even in the product development process on the basis of information collected in questionnaires. Face-to-face or phone contacts are not the only means of communication anymore. The Internet and other new communication media such as PDAs, WAP-wireless application protocol- (mobile phones, pagers, two-way radios, smartphones and communicators) and digital TV allow companies to interact with customers much more directly and in real time.

Although customization strategies are easier and cheaper with the available technology, the strategic and organizational decisions are more complex and expensive. A company must bring together the supply and demand sides of the market for successful customization. Managers face critical decisions about where and when to customize and how to integrate this strategy with other marketing strategies. Customization begins with the database. To compile a customer database, one needs to collect customer information, which is very costly. Money and staff resources available for the firm to do this have limits. Time is a major constraint. The value of the information gained has to be weighed against some estimate of the cost of its collection. Most direct marketers collect extensive household

## *Introduction*

information; however, there is a need to develop new methods to exploit this information fully to customize the product offerings or merchandizing strategies. In the next section, we briefly discuss marketing mix customization and customizability.

### **1.2 Marketing Mix Customization**

While customers are passive participants in traditional marketing, they are becoming active participants in customization through the processes of creating and marketing the product and service. Customers are more active in every stage of these processes through the Internet. The Internet makes it possible for customers to drive the process: to search for information they need to make choices, to create their own products and services, to set their own prices, and to self-select themselves into segment. Now, we explain how to customize five different instruments of marketing mix -- product, purchase price, communication, distribution, and after-sales support and cost-- either by the marketing firm or by consumers, with examples below and summarized in Table 1.1 (This table and section is mainly based on Logman, 1997). Some customization applications in marketing literature are illustrated in Table 1.2.

Table 1.1: Marketing mix customization and customizability options (Logman, 1997)

<i>Elements</i>	<i>By Company</i>	<i>By Customer</i>
Product	Offering enhanced and/or bundled products (to meet individual customer needs)	Offering final products with different options Offering a menu of product components (from which customers can select and design their final product)
Purchase price	Price discounting (dependent on sales volume, sales history, time of purchase) As a result of product customization	As a result of product customizability As a result of customers' bargaining power As a result of customers' decision timing
Communication	Using one-to-one communication tools (direct mail, sales force)	Offering a customizable interactive information network (such as the Internet)
Distribution	Offering multiple channel solutions (partly customizable)	Offering a customizable distribution network
After-Sales Support and Costs	Offering augmented product solutions (with single or bundled services) Using remote control systems	Offering do-it-yourself solutions As a result of product customizability (such as the way the product is used)

### **1.2.1 Customizing Product**

Customers can create final products from choice menus according to their needs, budgets and preferences. Now, many companies have websites, which allow online customization. Dell, for example, has a website that allows different hardware configurations for customers. Other examples are the customization of cars by GM and Chrysler, in the car industry, and custom-made jeans ([www.operand.com/portfolio/levis.php](http://www.operand.com/portfolio/levis.php)) in the clothing industry. Companies offer enhanced products (i.e. an enhanced product is a core product that has been differentiated by adding such tangible properties as features, styling, and quality) and/or bundled products (such as computer companies offering PCs with already installed software or printers) to meet individual customer needs. For this reason, they prefer to buy customizable products from suppliers and to adapt them or develop integrated solutions using modular systems (see Stremersch et al., 2003).

## *Introduction*

For instance, the Laboratory of Production Technologies of Siemens in Belgium uses integrated solutions to create products, which can be used in the production lines of different Siemens products.

### **1.2.2 Customizing Price**

Companies can customize their product price as a result of a customer's product customization, or by offering price discounts which are based on a customer's past purchase history, a customer's sales volume, time of purchase or product bundling. Customers can control prices through their bargaining power, which is possible by choosing the right moment to buy a product (such as waiting until the price drops) or searching for prices from different websites at different time points. Some websites, such as priceline.com, dealTime.com, and online auctions allow customers to customize purchase price. More companies, such as Chrysler or General Motors, allow their potential customers to design a product based on their own choices from available specifications, and calculate the price of the product using those specifications.

### **1.2.3 Customizing Communication**

Different information needs of customers, such as for new product versions, possible upgrades of old products, promotional or product information, call for customization methods. This customized information can be distributed to customers directly through direct mail or personal contacts, and through the Internet via websites or email. Internet advertising is the main tool for communication customization in the digital marketplace, since advertising messages can be rapidly distributed at very low cost, and are easy to

produce and distribute over the web and email. The customization of content, format, the educational component or entertainment power of the communication, mode of delivery, timing and place are becoming popular topics in marketing in recent years (see Table 1.2).

Table 1.2: Some customization applications in marketing literature

<i>Study</i>	<i>What</i>	<i>Method</i>
Rossi, McCulloch and Allenby (1996)	Customization of promotions Target couponing	Random coefficient choice model with individual level heterogeneity
Ansari, Essagaier and Kohli (2000)	Customization of offerings (Recommendation systems)	Hierarchical Bayes estimation
Gooley and Lattin (2000)	Customization of Marketing Messages Which content to present to whom	Multi-armed bandit problem approach maximizing response rate
Liechty, Ramaswamy & Cohen (2001)	Customization of communications Web-based information service	HB multivariate probit model
Raghu, Kannan, Rao & Whinston (2001)	Customization of communications	Information theory, segmentation, clustering techniques
Ansari and Mela (2003)	Customization of email-messages	Hierarchical Bayes (HB) estimation and combinatorial optimization
Bertsimas and Mersereau (2003)	Customization of marketing messages	Adaptive sampling
Toubia, Simester and Hauser (2003)	Customization of adaptive conjoint questionnaire	Polyhedral question design for partial profile conjoint. Analytical center estimation
Montgomery, Hosanagar, et al. (2004)	Designing a better shopbot Which stores to search, how long to wait, and which offers present to the user	Random utility model Decision approach
Zhang and Krishnamurthi (2004)	Customization of promotions When to promote how much to whom	Incidence-choice and quantity model Optimization
Zhang and Wedel (2004)	Customization of Promotions: comparison of market, segment-based, personalized	Incidence-choice and quantity model Profit optimization

### **1.2.4 Customizing Distribution**

Customers now have more freedom in selecting the logistics and the methods of distribution to meet their needs. New distribution strategies are developed with the increasing usage of the Internet. Customers can determine where, when and how they want goods to be delivered, and in which manner. Amazon.com, for instance, offers three different delivery

## *Introduction*

timings with different prices. In electronic shopping, customers can continuously monitor and adjust orders, schedule delivery, places of distribution and how they want goods to be delivered. Companies prefer to use multiple channels for distribution flexibility depending on the customer's product knowledge, service needs, and future price sensitivity.

### **1.2.5 Customizing After-Sales Support and Costs**

Customers can choose do-it-yourself solutions, which are offered by the company, and buy customized products, which come with a customizable information network for after-sales support. Companies generally use remote-control systems for after-sales support. The Internet is one of the best tools for customized after-sales support. Especially in the computer industry, companies offer customized augmented solutions which include product, training, service or logistics offers, such as product maintenance, replacement, and so on. For example, some software companies use this method for updating software applications or fixing problems online.

### **1.3 Problem Delineation:**

This thesis deals with aspects of these two key components of the customization process: 1) efficient customized data collection and 2) customization of marketing mix across multiple product categories. Both aspects involve stochastic modeling of consumer behavior/response, and optimal decision making based on the first process. Customization of marketing actions given limited information depends on inferences and the characterization of the level of uncertainty in these inferences. From this

perspective, Bayesian techniques are the most suitable tool for the problems anchored in this thesis. The Bayesian framework enables an elegant integration of response models and decision making which incorporates the uncertainty of model estimation in the decision framework. Such decisions of “what to ask whom” and “what to promote to whom” are at the core of this thesis.

We have seen an enormous increase in the use of Bayesian techniques in marketing in the past decade. The main reason behind this is that Bayesian methods are particularly appropriate to the decision orientation of marketing problems, and further, they ideally suit a wide range of marketing data and decision processes. Bayesian data analysis has the ability to handle many different types of response variables in the same analysis. Since marketing data are often lumpy and not very well-suited for making standard distributional assumptions, Bayesian methods have come to play a critical role in marketing models. Marketing data are also sparse at the individual level in general. While we need large samples in frequentist methods, for approximations of standard errors, we use posterior distributions in Bayesian inference, which enables accurate inference for all parameters and all sample sizes. That is, all Bayesian results are exact in finite samples because the distributions are derived conditional on the observed sample of data. However, many classical theory results depend on asymptotics and are only approximations for the observed sample of data. Missing responses in Bayesian analysis are easily modeled as latent variables in a manner that uses the information contained in observed data.

## *Introduction*

Marketing models often include latent variables, especially in consumer behavior and decision making problems. While frequentist methods allow for few latent variables (except for structural equation modeling) due to estimation difficulty, Bayesian models enable many latent variables to be included in a relatively straightforward fashion. Most data for marketing research is generated according to a hierarchical process, and again, such hierarchical models are easily implemented in Bayesian analysis. Hypothesis testing is different for the two approaches: While Bayesians measure the data's support for the hypothesis, classical statisticians measure the hypothesis' support for the data. Detailed explanations on the advantages of Bayesian data analysis in marketing can be found in Elrod (2005), and Bayesian statistics applications in marketing can be found in Rossi and Allenby (2003).

Bayesian statistics have been criticized by classical statisticians for the subjective prior information used. The prior information however can also be "objective." Practically, prior information may in fact improve decision making (Berger 1985). In marketing, prior information is readily available from huge databases which are collected by market research companies. Some researches in marketing use prior information in their estimations such as from experts (e.g. Sandor and Wedel 2001, Popkowski and Sinha 2005), or prior theory (e.g. Montgomery and Rossi 1999), or other datasets (e.g. Lenk and Rao 1990, Putler et al. 1996, Kamakura and Wedel 1997, Wedel and Pieters 2000, Ansari et al. 2000, Ter Hofstede et al. 2002).

## **1.4 Motivation of Essay 1:**

In the first essay, we study how to design optimal split questionnaires, which helps to collect better quality data faster and cheaper. Usage of the Internet is doubling every year<sup>1</sup>. This rapid growth of the Internet creates an opportunity for conducting online marketing research. By some estimates, about 60% of the population of the United States and the European Union has Internet access. This widespread adoption of the Internet makes a large cross-section of the population accessible and ensures that information on the needs and preferences of a substantial population of the consumers can be obtained online. In 1995, some of the first articles were published comparing email with postal surveys. For example, Mehta and Sivadas (1995) showed that email could generate high response rates similar to postal surveys.

Moreover, high levels of product customization need extensive profiling and customization tools to identify and target individual customers, based on a combination of demographics, attitudes and past interactions. Growing numbers of organizations and companies need to use more sophisticated means to get information on their Web site visitors. For these companies, online questionnaires can be a tool to link future customers to specific products and services. Companies can utilize analyses of consumer interests on Web sites. A recent survey among companies by WIT inc., a

---

<sup>1</sup> [www.virtualsurveys.com/news/papers/paper\\_9.asp](http://www.virtualsurveys.com/news/papers/paper_9.asp)

## *Introduction*

Web services provider, found that 55 percent of respondents in Michigan plan to upgrade customer relations on their sites in 2004.

Campbell-Ewald Digital, a Warren-based advertising and marketing company, is one of these companies, and according to its senior vice president/creative director Harvey Zuppke, more companies are turning to cultural anthropologists and psychologists to develop online surveys that will produce profiles of potential customers and a broader picture of their lifestyles, in efforts to build a relationship with their customers. All of the market research companies' clients use online surveys to learn more about their customers and how they interact with these sites. For instance, a Chevy Malibu Internet site by Campbell-Ewald questions visitors about their driving habits and what they value most in a car. After analyzing their answers, the Web site provides information about the car's features that should be most appealing to them. Many more companies use interactive questionnaires to help customers find the right product and help the company determine its customer base.

Developing the questions can be a complicated and time-consuming process, and long questionnaires that inquire about potential customers' lifestyles, attitudes, needs and past behavior may cause problems of attrition, nonresponse, fatigue and boredom of potential customers, and may not even be feasible on the Internet. Any efforts to improve the quality of the data will increase the effectiveness of market actions based on it. From this perspective, split questionnaire survey designs (which are not only useful in online surveys, but also for paper or phone surveys) help

market researchers to provide faster, cheaper and efficient ways of collecting data about customers.

The increasing usage of online marketing research needs more advanced methods to collect data, and from this respect the need for better questionnaire designs is increasing. Split questionnaires, adaptive questionnaires and individual level customized questionnaires have great potential for use in online surveys. In split questionnaire survey design, the original questionnaire is divided into sub-components and subjects respond to a randomly selected set of components only. Finding an optimal design for a split questionnaire involves finding the configuration of question sets (i.e. those questions given to one respondent, or a “split”) such that a minimum amount of information is lost as compared to the complete questionnaire. Some ad-hoc splitting strategies often used in practice may depend on the purpose and the contents of the survey, contextual placement of certain items, and the partial correlation coefficients of the items (Raghunathan and Grizzle, 1995). We suggest, in line with previous practice in marketing research, to utilize the natural structure of the questionnaire, in which questions are placed in blocks. Mostly, several questions measuring, for example, one particular attitudinal or lifestyle trait are administered as a group or block. We use this block-structure to generate split-questionnaire designs in two different ways: selecting entire blocks of questions, which we call a “between-block design”, or selecting questions in each block, which we call a “within-block design”. In the between-block design, a “split” comprises of the allocation of selected

## *Introduction*

blocks of questions and respondents answer all questions in these blocks; in the within-block design, a split comprises of sets of selected questions in each of the blocks and respondents answer only those questions in each block. For the first method, our research problem then simplifies to how these blocks should be administered to respondents in an optimal way. On the other hand, for the within-block design, our research problem is how to choose questions in each block optimally. After we generate optimal split questionnaires, we administer these different versions of the questionnaires and finally we multiple impute data with the Gibbs sampler for the missing responses using information from other subjects that responded to the missing parts.

In the questionnaire design area there are several possibilities for custody of a good design: The first is to reduce questionnaire length by dropping out uninformative questions. Factor analysis can be used for this approach. The second is to find user profiles from the sample data so that future users can be classified according to those profiles with classification methods (especially discriminant analysis) and offered different versions of the questionnaire (Zhang and Fang 2003, Haaland et al. 1979, Brockett et al. 1981). We compare our approach to these two alternatives. In Chapter 3, we detail how to design optimal split questionnaires. Our approach --split questionnaire design-- differs from those methods in two ways. First, instead of dropping some questions from the questionnaire, we use all questions, only different people respond to different parts of the questionnaire. Second, instead of classifying subjects, we generate different versions of the questionnaire based on prior information.

## **1.5 Motivation of Essay 2:**

Companies have become increasingly interested in customization possibilities of interactive media as a result of significant advances in technology. Interactive media allows the marketer to identify the consumer and characteristics of the consumer, decide on the marketing message in real time and capture response to marketing communications. For instance, e-commerce sites such as amazon.com and dell.com can customize content (e.g., information, digital products such as software, advertising, promotions, recommendations...) to increase purchases. Nowadays, increasing numbers of companies develop customized and targeted online programs such as customized ads, websites, email-messages, customized sales-promotions to loyalty card users, customized electronic coupons, etc. Targeting and customization issues have long been of interest in marketing. In the previous sections, we have mentioned customization and customizability options for marketing mix instruments. In the second essay, we focus on the customization problem of how to develop promotion designs across multiple product categories simultaneously.

The dynamic nature of the Internet (and other interactive media) is particularly suited to offer promotions to individual customers “on the fly” to guide their decisions by using information from their previous decisions. Hence, delivering promotions individually via email or the web, one of the main interests of online customization, is becoming a more important subject. Specifically, online grocery stores such as Peapod (peapod.com)

## *Introduction*

and NetGrocer (shop.netgrocer.com) possess the technological potential to customize the grocery shopping process. Currently, Peapod allows customers to create personal lists, such as frequently purchased products, products purchased for weekend parties, and products for special occasions (e.g., Thanksgiving) for its customer. Using this service, customers can reduce their shopping time, eliminate product categories of no interest to them, and keep checking totals of purchases so that they can spend within their budgets. Peapod also customizes the shopping experience by helping customers to list the items available in their pantry and refrigerator, and then suggesting recipes where these items can be used. These companies should fully utilize their technology and explore the potential for offering customized promotions. An example of using a customized promotion program is CVS Pharmacy. They use loyalty cards to offer different sales-promotions for low-tier, middle-tier, and top-tier customers. Additionally, they use targeted health mailings with segment level content and customized offers. They also target offers at the register using previous category purchase histories.

Since the decision of which items to promote to whom is very important, we consider the development of a customization method by focusing on the selection of target categories to be promoted from multiple product categories in an online shopping environment. Our method can be applied to different promotion programs such as individual specific e-coupon or point-of-purchase coupon distribution, and individual specific advertising. Personalized advertising and promotions are pervasive in a wide range of industries including services such as banking, telephony, insurance, durable goods such as autos, and a vast range of products sold in

supermarkets and drugstores. Currently, electronic coupons are issued by companies based on customer information in a way that does not depend on the (multivariate) relationships in purchase expenditures between categories. Our approach aims to obtain cross-category information and use this information in customizing coupon programs. In particular, web pages of specialized online coupon companies (e.g., couponmountain.com, coolsaving.com, couponcabin.com, and addcoupon.com) show a certain number of coupon offers and our purpose is to select the most suitable (profitable) categories to offer to individuals to minimize the search effort, as well as maximize the retailer revenue. The e-coupon is a short piece of text that can carry a commercial message, including price and availability of product in question. Electronic distribution of coupons has become more widespread under programs such as Catalina Marketing Incorporated's (CMI) Checkout Coupon and Frequent Shopper schemes (in-store coupon distribution), in which households receive coupons offering discounts through the Internet (see [valuepage.com/Entry.pst](http://valuepage.com/Entry.pst)). According to the Association of Coupon Professionals, Internet-delivered coupons, although still a controversial topic in the industry, saw a five-fold increase in distribution as entrepreneurial marketers sought better ways to target and deliver effective incentives ([couponpros.org](http://couponpros.org)).

There are three key components in this essay. The first one is multicategory modeling. We fit a hierarchical Bayes type-2 multivariate tobit model which allows us to estimate individual and average level consumer preferences, cross-category incidence, expenditure and incidence-

## *Introduction*

expenditure correlations using purchase incidence and expenditure data. Multicategory models are particularly appealing in our context because (online) retailers aim to maximize store profits by jointly coordinating marketing activities across product categories. Manufacturers that sell products in multiple categories may also benefit from these models, since they can use this information in production, price setting or for product bundling. Service provider firms may be interested in undertaking cross-selling initiatives across product categories. The second concept is individual level heterogeneity. We include individual level heterogeneity in the coefficients of marketing activities for each individual consumer. Marketing models need to consider individual heterogeneity, since consumers may react differently to the marketing activities (marketing mix, such as price and promotion) and these differences between individuals form the very basis of customization. The last concept is optimal decision making. We estimate expected expenditures of each customer for each category and select the optimal combination of five categories to offer from among many. We consider parameter and estimation uncertainty in our estimation using the Bayesian decision framework. Importantly, we use the Bayesian approach to addressing these three concepts, since Bayesian statistics optimally investigate inference, estimation and decision problems of marketing.

## **1.6 Outline of the Dissertation**

This thesis contains two essays on dealing with how to more efficiently collect data and how to customize online promotion offers across multiple

categories. In the first essay, we introduce a method to design split questionnaires to collect data more efficiently, i.e. faster, cheaper and with better quality, using experimental design techniques. In the second essay, we develop a customization approach and propose a method of optimizing the selection of categories to promote based on the Bayesian decision framework, using online grocery retail data. The Bayesian decision framework is used in both essays.

In Chapter 2, we discuss Bayesian statistics for inference and decision problems in marketing. After giving some insights for Bayesian statistics, we explain in detail why Bayesian methods are commonly accepted by the marketing community and discuss some advantages of it for marketing. We focus on the Bayesian approach for marketing decision problems. We explain briefly some Bayesian estimation algorithms used in both essays.

In Chapter 3, we focus on a split questionnaire survey design. This involves subsets of subjects responding to different parts of the questionnaire instead of the whole. Chapter 3 deals with the problem of constructing an optimal split questionnaire design, which means asking fewer questions per subject to obtain the most information. Split questionnaire design results in data missing by design. Our purpose in this chapter is to develop a method, using experimental design techniques, to select the best allocations of blocks of questions and question allocations in each block for splits (i.e. to generate different versions of split questionnaires) to maximize information. We reduce respondent burden with this method by asking fewer questions per subject. We explain the

## *Introduction*

proposed optimal split questionnaire method, which is based on prior information, and optimization by a design generating algorithm -the modified Federov algorithm- to find the optimal design from all possible designs. We explain how to construct identified split questionnaire designs, and how to impute the missing data with the Gibbs sampler. We also present empirical and simulated data results to illustrate the statistical efficiency of this method. This chapter is based on Adiguzel and Wedel (2004).

Respondent burden is related to the time and the effort a respondent has to expend to complete a questionnaire. Time and effort are a function of the length and the nature of the individual items in a questionnaire. Therefore, it is reasonable to expect a degree of correlation between respondent burden and quality of the data. A reduction in respondent burden may also have a positive impact on reducing item nonresponse rates. We investigate behavioral effects of using split questionnaires and illustrate these effects on data quality in Chapter 3 in a field study.

In Chapter 4, we define the problem of promotion customization. In this chapter, we provide a literature review of multivariate category applications in the marketing literature. For efficient customization, we need individual level customer information, and for that purpose we develop a model to analyze purchase incidence and expenditures of multiple categories. The model is a hierarchical Bayes multivariate type-2 tobit model and is estimated with the Gibbs sampling. Based on that, we develop an optimization algorithm to choose the optimal combination of categories from among many to promote for each customer. Our approach maximizes each

consumer's total expenditures among all categories involved using the Bayesian decision framework. The approach is based on design generating algorithms used in experimental design literature (i.e. modified Federov algorithm) to solve this combinatorial optimization problem. We investigate our model and its performance on synthetic data, and give the applications and results of this problem.

Finally, in Chapter 5, we present conclusions and discussion of substantive issues in these two essays, and describe the possible venues for future extensions.



# Chapter 2

## Decision Making under Uncertainty

### 2.1 Introduction

Bayesian decision analysis is a powerful tool to many professionals: market researchers, operations researchers, statisticians, businessmen, economists, engineers, psychologists, computer scientists and those in other fields where prediction and decision making must follow from statistical analysis. The Bayesian statistical tradition provides a formalized way of learning about the parameters of a statistical model from data and originated in 1763, with the theorem formulated by Reverend Thomas Bayes. The Bayesian paradigm has received tremendous popularity in marketing, since it affords an exceedingly flexible and robust framework for developing and estimating statistical models that facilitate realistic description of marketing data. Such models may include latent variables, missing data, mixed outcome data, heterogeneity of coefficients, and more. In particular, in Chapter 3 the model formulated for survey data includes missing values (due to the design of questionnaires) and possibly mixed outcome variables (i.e. rating scales, binary pick-any items, categorical demographic variables, etc.). In Chapter 4, we develop a model with heterogeneity of coefficients (individual level price and promotion sensitivities) and mixed outcomes (the incidence of a category and the expenditure on it). A basic paradigm in marketing is the notion that customers differ in their preferences, needs and choices, and that firms need to take such differences into account in determining optimal marketing actions. Rossi and Allenby (2003) postulate that statistical analysis of

marketing data is comprised of three components: within-unit behavior and across unit heterogeneity in that behavior (where unit can be a consumer, a household, or an organization), and action -the solution to the marketing decision problem that recognizes these previous components. Marketing data typically is comprised of many heterogeneous units, often with only limited information on each unit. The statistical problems associated with accommodating heterogeneity of consumers in statistical models and the subsequent management decisions have propelled the use of Bayesian statistical methodology.

It is fair to say that in marketing, the Bayesian paradigm is now the dominant paradigm for inference and decision making in such diverse areas as pricing, new product development, promotions, conjoint analysis and the design of conjoint experiments and --of particular interest to this thesis-- customization of marketing instruments to individual consumers. Some applications of Bayesian approach are in pricing (Montgomery, 1997, Montgomery et al., 1999, Kalyanam et al., 1998, Kalyanam, 1996); in new product development (Neelamegram et al., 1999, Lenk et al., 1990, Allenby et al., 1995, Talukdar et al., 2002, Michalek et al. 2005); in promotions (Blattberg et al., 1991, Boatwright et al., 1999); in conjoint analysis (Allenby et al., 1995, Andrews et al., 2002, Marshall et al., 2002, Otter et al., 2003, Bradlow et al., 2004); design of conjoint experiments (Sandor et al., 2001, Lenk et al., 1996); consumer demand modeling (Allenby et al., 1998, Kim et al., 2002); advertising (Wedel et al., 2000); in customization (Ansari et al., 2003, Liechty et al., 2001); and Internet applications, such as

### *Decision Making under Uncertainty*

recommendation engines, web-browsing behavior etc. (Ansari et al., 2000, Bradlow et al., 2000, Sismeiro et al., 2004, Rahul et al., 2004). See also the review of Rossi and Allenby (2003).

Rossi, McCulloch and Allenby state in their seminal paper in 1996: “Any successful customization approach must deal directly with the problem of partial information and take parameter uncertainty into account in the decision problem”. In this thesis, we will apply the Bayesian paradigm to facilitate decisions on “what to ask whom” in the construction split questionnaires and on “what to promote to whom” in designing optimal promotional plans, in Chapters 3 and 4 of this thesis. In this chapter, we discuss the Bayesian paradigm, advantages of this approach on a classical approach, and Bayesian decision making that will be used commonly later. In this chapter, we follow Rossi and Allenby (2003), Rossi, McCulloch and Allenby (2005) and Lenk and Wedel (2001).

Bayesian methods propose the optimal way to make consistent decisions in the face of uncertainty. The reason behind this is that Bayesian statistics seek to optimally combine information from two sources: the information that we have or believe at the start of the research and the information in the observed data. Bayes theorem provides the mechanism to combine these both sources of information into a single set of updated information (i.e. the posterior distribution) of the quantities of interest.

In statistics, we quantify uncertainty in observable scientific data through probability distributions, which depend on unknown quantities, called parameters. In the Bayesian paradigm, current knowledge before data analysis is represented with a prior distribution, and updated

knowledge based on available data is represented with a posterior distribution. Explicitly, current knowledge about the model parameters is expressed by placing a probability distribution on the parameters, called the "prior distribution", often written as  $p(\theta)$ . This prior distribution can take on a variety of well known forms, such as the Normal, Binomial, Bernoulli, Poisson, and Gamma distributions, but also multivariate distributions such as the Multivariate Normal, Multinomial, Dirichlet and Wishart (see Casella and Berger, 1990, for an overview of these distributions and specific details). When new data  $y$  becomes available, the information they contain regarding the model parameters is expressed in the "likelihood," which is proportional to the distribution of the observed data given the model parameters, and written as  $p(y|\theta)$ . Thus, the likelihood requires the specification of a distributional form for the data, as a function of the unknown parameters  $\theta$ . The information in the data as contained in the likelihood is then combined with the prior to produce an updated probability distribution called the "posterior distribution,"  $p(\theta, y)$  on which all Bayesian inference is based. Bayes' theorem illustrates how this update is done mathematically and shows that the posterior is proportional to the prior times the likelihood,

$$p(\theta | y) = \frac{p(\theta) \times p(y | \theta)}{\int p(\theta) \times p(y | \theta) d\theta} \propto p(\theta) \times p(y | \theta) \quad (2.1)$$

This posterior distribution captures the uncertainty in the parameters after the data has been observed, but can take complex forms for realistic

### *Decision Making under Uncertainty*

models. A major breakthrough in the application of the Bayesian paradigm was the realization that in many cases the expressions for the joint posterior distribution of multiple parameters can be factored into simpler expressions that can be recursively sampled from, using so called Markov Chain Monte Carlo methods (MCMC, Geman & Geman 1984, Gelfand & Smith 1990). Markov Chain Monte Carlo (MCMC) simulation has enabled the estimation of complex models that are nearly impossible or very difficult to estimate with classical methods (Gelfand and Smith 1990, Smith and Roberts 1993, Gilks et al. 1996). This holds in particular for hierarchical Bayes models, which have received much popularity in marketing because of the importance of individual differences in a wide variety of models, and the need to investigate factors that influence those. A main theoretical advantage of the Bayesian framework is that while we examine the probability of the data given a model (hypothesis) in frequentist statistics, we examine the posterior probability of a model --or its parameters-- given the data in Bayesian statistics. This, for example, allows for accurate inference in small samples.

In Markov Chain Monte Carlo algorithms, one generates a large sample of independent draws from the posterior distribution, and each draw is conditional on the previous one. MCMC is a Monte Carlo integration using Markov chains. The transition probabilities between sample values are only a function of the most recent value, which is why the technique is referred to as a Markov Chain. The Monte Carlo term comes from the Monte Carlo integration in which we draw samples from the distribution, and then calculate sample averages to approximate expectations. The Gibbs sampler (Geman and Geman 1984) is the most commonly used Markov

Chain Monte Carlo method and widely applicable to various Bayesian problems. We used the Gibbs sampler to impute missing data in Chapter 3 based on a survey response model, and to estimate the parameters of a hierarchical Bayes multivariate type-2 tobit model in Chapter 4.

Gibbs sampling simply means sampling from the full conditional distributions. Suppose that there is a random vector  $Y$  which consists of  $J$  subvectors,  $Y=(Y_1, Y_2, \dots, Y_J)$ , and the joint distribution of  $Y$ ,  $P(Y)$ . We iteratively draw from the conditional distribution of each subvector given all the others in Gibbs sampling. We represent this given the value of  $Y$  at each step  $t$ ,

$$\begin{aligned} Y_1^{(t+1)} &\sim P(Y_1 | Y_2^{(t)}, Y_3^{(t)}, \dots, Y_J^{(t)}) \\ Y_2^{(t+1)} &\sim P(Y_2 | Y_1^{(t+1)}, Y_3^{(t)}, \dots, Y_J^{(t)}) \\ &\vdots \\ Y_J^{(t+1)} &\sim P(Y_J | Y_1^{(t+1)}, Y_2^{(t+1)}, \dots, Y_{J-1}^{(t+1)}) \end{aligned} \tag{2.2}$$

Gibbs sampling is closely related to data augmentation (Tanner and Wong, 1987). Data augmentation refers to methods for constructing iterative algorithms via introduction of unobserved data or latent variables. Many models in marketing which contain latent variables use this method, including limited dependent variable models (like choice or censored regression models), state space, or common factor models, and models with heterogeneity. Gibbs sampling is especially well suited to coping with

incomplete information and we illustrated the application of Gibbs sampling on missing data in Chapter 3.

Loss functions are used to estimate parameters in the Bayesian approach and to make decisions. A loss function measures the loss caused by an estimation error or decision error. Estimation is a special case in decision-making, and the goal is to choose the estimator which minimizes the expected loss. In the case of estimation problems, the loss function is a function of the parameter estimate and the true (unknown) parameter value. Common choices of loss functions are quadratic loss, absolute loss and zero-one loss, which are represented below, respectively.

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2 \quad (2.3)$$

$$L(\hat{\theta}, \theta) = |\hat{\theta} - \theta| \quad (2.4)$$

$$\begin{aligned} L(\hat{\theta}, \theta) &= 0 & \text{if } \hat{\theta} = \theta \\ &= c & \text{if } \hat{\theta} \neq \theta \end{aligned} \quad (2.5)$$

While the optimal Bayesian point estimate is the mean of the posterior distribution when choosing a quadratic loss function, the median of the posterior distribution is the point estimate when choosing the absolute loss function. The mode of the posterior distribution is the point estimate in the case of a zero-one loss function. Although the choice of loss function depends on the problem, the most commonly used is the quadratic loss function, which uses all the information in the posterior distribution. We will focus on the case of decision problems later.

We briefly discuss a few issues of the Bayesian approach, in particular the choice of the prior distribution. There are two kinds of prior information: objective and subjective (see more information on objective and subjective priors in Press, 2003). In the classical approach, all modeling assumptions are actually a kind of prior information, and generally the underlying assumptions can remain hidden, whereas the researcher's prior beliefs are expressed as prior information in the Bayesian approach. Prior beliefs can be obtained from experts, prior studies, theories or other data sets. Prior distributions are intrinsically subjective (everyone's prior information is different). Subjective Bayesian statistics, firmly rooted in probability theory (De Finetti, 1970), proposes that a model reflects a researcher's belief about a phenomenon and that people can and should conceive of uncertainty about events as subjective probabilities (Savage, 1954). This at the same time raises a reservation some have about the Bayesian approach: posterior predictive inferences are sensitive to the choice of the prior, and so are decisions made based on the model inferences. Many "pragmatic Bayesians" (See Lenk and Wedel, 2001), predominantly concerned with the flexibility of model construction that Bayesian statistics now afford through MCMC methodology, therefore choose non-informative priors for their model. Many statistical models formulated in the Bayesian framework, are therefore based on non-informative or weakly informative priors, which minimize the influence of prior assumptions on posterior inference. Although many statisticians see the subjective priors as a fundamental drawback of the Bayesian approach, this is inescapable, and

### *Decision Making under Uncertainty*

frequentist methods (classical statistical theory) also entail subjective choices. However, using prior information well may in fact improve decision-making (Berger 1985). Certainly, the investigation of the sensitivity of the predictive distribution to the specification of the prior is critical. In case subjective prior distributions for the model parameters can be assessed, we may need to elicit priors from consumers, decision makers or other subject-matter experts. Allenby et al. (1995) state that incorporating expected prior ordinal information of attributes into conjoint analysis improves the estimation. Sandor and Wedel (2001) illustrate that this approach is attractive in the design of choice experiments. Popkowski and Sinha (2005) propose a method to facilitate the subjective information from a modeler or manager into a choice model and illustrate the improvement on the marketing strategy (decisions). Wolfson (1995) and Chaloner (1996) provide an overview of the various philosophies of elicitation based on the ways people think about and update probabilistic statements. The use of loss functions to make optimal decisions concerning settings of control variables allows the statistical process to be customized to fit the particular application in question. Emphasis in Bayesian marketing is now shifting from inference towards the decision problem. Although quickly improving in quality, the models may not realize their full potential in decision making until put into the framework of the decision-making process. Better decisions require better procedures to extract information from data and incorporate that in the marketing decision problem, and the Bayesian paradigm is optimally suited for that (Lenk and Wedel, 2001). I provide two examples of this in Chapters 3 and 4 of this thesis.

## **2.2 Bayesian Analysis and Marketing Decisions**

In the previous section, we discussed the Bayesian approach for statistical inference and here we discuss the Bayesian approach for decision-making. The main purpose of decision theory is to develop techniques and methods that facilitate making decisions in an optimal way. The standard estimation problems of statistical inference or testing hypotheses can be formulated as decision problems (see Cyert and DeGroot, 1987). In this section, we compare “Bayesian approach” in decision theory to the naïve approach, “plug-in.”

Most of the model parameters on which marketing action decisions are based, are unknown random variables, and we need to choose the optimal value of deterministic control variables, the effects of which depend on those random parameter values. Thus in the decision process, the parameter estimation or model uncertainty must be considered. The Bayesian approach merges all available prior and observed data information to estimate the parameters that are the basis of the optimization problem. Levels for the control variables can then be found that maximize or minimize the expected value of an objective function (Berger, 1985). Bayesian decision theory is very appropriate for marketing decision problems. The reasons are 1) subjective prior probabilities from economic theories, experts or managers can be easily used to express pre-existing information as prior information that improve the estimation, 2) it entails careful modeling of the data structure, checking and allowance for

### *Decision Making under Uncertainty*

uncertainty in model assumptions, 3) formulating a set of possible decisions can be easy, and 4) utility functions are used to express how the value of each alternative decision is affected by the unknown model parameters. The objective in decision-making is to choose an action that minimizes the expected value of the loss function with respect to the posterior distribution, if data are available. However, if data are not available, the expected loss should be minimized with respect to the prior distribution<sup>2</sup> (see Sandor and Wedel, 2001, on designing conjoint experiments).

The main goal of Bayesian decision theory is to minimize the expected loss of a decision or minimize the expected risk. To do this, Bayesian decision theory leads to an optimal decision considering the expected loss of all possible values of the random parameters values by weighting those by the probability of their occurrence (i.e. posterior distribution). A loss function and the posterior distribution are the two main components in decision theory and will be explained below. In Chapter 3, we develop a model for survey data, with a missing data structure that is due to the design of the questionnaire. Then, we formulate a loss function that captures how far a data-structure with missing data is from a true, complete dataset. The decisions are then what questions to pose to which respondents, to minimize the loss. In Chapter 4, the decision is what categories to promote to which customers, based on a joint model of category incidence and expenditure sensitivity to prices and promotions.

---

<sup>2</sup> The mean of the posterior distribution is the Bayes estimator with respect to the quadratic loss function. If no data are available, the posterior distribution reduces to the prior distribution, and the Bayes estimator becomes the mean of the prior distribution (see details Press, 2003).

The loss function is related to the incremental revenue obtained from promoting a specific set of categories.

Suppose a manager has two or more alternative actions, “a”, for any decision. Each action “a” has some potential loss that will depend on the parameters,  $\theta$ , and this relationship is expressed through the loss function,  $L(a, \theta)$ . For example, a marketing manager has to decide what category to promote for which consumer (Chapter 4). The different possible promotions are the alternative actions. The state of nature is the likely demand for every possible set of promotions, and revenue is the loss function that relates promotion to demand. We can calculate the expected revenue/profit (or loss) for every possible combination of allocation of promotions and demand. Management objectives are specified as a function of model predictions (and/or parameters), for example the predicted revenue arising from promoting a specific category, and the expected consequences of any particular management action (i.e. every possible set of promotions) are calculated by integrating over the uncertainty in both model parameters and model predictions. If we express this mathematically, the optimal decision maker chooses the action so as to minimize expected loss, where the expectation is taken with respect to the posterior distribution (Rossi and Allenby 2003)

$$\min_a E[L(a)] = \int L(a, \theta) p(\theta | \text{data}) d\theta \quad (2.6)$$

This integral can be evaluated either numerically or by Monte Carlo integration. The explicit incorporation of the posterior distribution of the

### *Decision Making under Uncertainty*

random variables (which includes prior information) makes the decision theory approach Bayesian. In decision-making, uncertainty means that the outcome of a decision maker's action is not exactly predictable because of the unknown parameter values or random error terms. The traditional approach only assumes a probability distribution for error terms but not for the unknown parameter values. Such an approach does not permit the choice of an optimal decision, which reflects estimation uncertainty. However, in the Bayesian approach, prior distributions are assigned for unknown values of the all parameters in the decision problem.

Consider the more complex marketing decision problem in which we have explanatory variables  $x$ , consisting of some control variables  $x_c$  and the remaining explanatory variables,  $x_f$ . For example, we may have promotion as a (0/1) control variable, and price as another explanatory variable. We have a probability distribution  $p(y|x,\theta)$  which represents how the dependent variable (outcome) is related to explanatory variables. The decision maker wants to choose  $x_c$  to maximize the expected profits or revenue where the expectation is taken over the distribution of the outcome variable. In a fully Bayesian decision approach, this expectation must be taken with respect to the posterior distribution of  $\theta$  and the predictive conditional distribution of  $p(y|x_c, x_f)$  and expressed by Rossi and Allenby (2003) as:

$$\begin{aligned}\pi^*(x_c | x_f) &= E_\theta [E_{y|\theta} [\pi(y | x_c)]] \\ &= E_\theta \left[ \int \pi(y | x_c) p(y | x_c, x_f, \theta) dy \right] \\ &= E_\theta [\bar{\pi}(x_c | x_f, \theta)]\end{aligned}\tag{2.7}$$

Sometimes, a “plug-in” method is used, where the (maximum likelihood or posterior mode) estimates of any unknown random parameters are inserted into the optimization problem and then the decision problem is solved. In other words, the plug-in approach disregards the uncertainty and assumes that parameters are estimated perfectly. However, the certainty equivalence theorem<sup>3</sup> demonstrates that if and only if the posterior distribution of the parameters is normal and the objective function is linear-quadratic in the unknown parameters the plug-in leads to the same solution as a Bayesian approach (Dorfman, 1997). In Chapter 3, we rely on that result when we use a plug-in estimator to compute the optimal questionnaire design, in the case where all measure variables are normal. However, we do not always have symmetric posterior distributions because of prior information or particular distributional assumptions, and the objective function is not always quadratic. Most of the time objective functions in marketing are based on expectations of the explicit function of profits, and these profit functions are generally not linear. In such cases, taking a Bayesian approach to solving for the optimal control will produce a different answer from the standard method. Expressing this mathematically (Rossi and Allenby, 2003):

$$\pi^*(x_c) = E_{\theta|y}[\bar{\pi}(x_c | \theta)] \neq \bar{\pi}(x_c | \hat{\theta} = E_{\theta|y}[\theta]) \quad (2.8)$$

---

<sup>3</sup> If the loss function is quadratic and the constraining model is linear and stochastic only by additive random disturbances that are independent of the instruments and whose expected values are zero, then the optimal values of the instruments are the same as if there were no uncertainty (Theil, 1954).

### *Decision Making under Uncertainty*

This is the approach we take in Chapter 4 when deriving optimal individual level promotion allocations. Bayesian methods are ideal for cases where we have prior information that shapes posterior distributions and for problems where risk is important and the correct objective function is not quadratic, but skewed. For this reason, Bayesian decision theory is often referred to as decision-making under estimation risk, and it relies on incorporating the uncertainty from estimation process into an optimal decision.





## Chapter 3

# Split Questionnaire Design

### 3.1 Introduction

Market researchers have traditionally collected consumer information on preferences, attitudes, consumption contexts and lifestyles, by means of often very long questionnaires. In doing so, they need to make tradeoffs between reasonable survey length and the value and quality of additional information. Questionnaire length is a concern since it affects the quality of the data collected in several ways (Berdie, 1989). Long questionnaires lead to higher non-response, item non-response and early break-off rates. They also cause an increase in the use of undesired response styles, increased time to collect the data, and respondent fatigue and boredom. It has been reported that survey respondents become fatigued and irritable when questioned for more than twenty minutes. Many studies indicate that longer questionnaires have lower response rates than shorter ones (Adams and Gale, 1982; Bean and Roszkowski, 1995; Dillman, 1991; Dillman, Sinclair, and Clark, 1993; Heberline and Baumgartner, 1978; Roszkowski and Bean, 1990).

#### 3.1.1 Motivation

We propose a method to design split questionnaire surveys as an effective tool to reduce respondent burden without sacrificing the inferential content of the data. Although Good (1969, 1970) already called for the development of split questionnaire methods to collect survey data more efficiently, in the following thirty-five years, no systematic research on how to best design

split questionnaires seems to have been done. Two decades ago, Herzog and Bachman (1981) advised that a researcher who needs to use a long questionnaire might be well advised to split the material into at least two parts and administer those parts in different orders to different random subsets of the sample. In their split questionnaire survey design, the original questionnaire is divided into sub-components, and subjects only respond to a randomly selected subset of components. A similar idea of designing randomly split questionnaires is applied in what has been called "time sampling". Here, questions are administered in a randomly rotated fashion to different parts of the panel in different episodes (Sikkel and Hoogendoorn, 1995). Incomplete designs in educational testing are based on a similar approach. In test construction, the researcher administers subsets of the total available item pool to the available subjects. The matrix sampling design (Shoemaker, 1973; Thayer, 1983), in which a test instrument is divided in sections, and groups of sections are administered to subjects in a randomized fashion, is used for that purpose.

Each of these previous studies has thus used a randomization approach to design split questionnaires. The important question that remains is how to optimally split the questionnaire such that the least information is lost. Currently, no methods have been published to address that problem, and here lies the contribution of this chapter. Raghunathan and Grizzle (1995) mention that ad-hoc splitting strategies may depend on the purpose and the contents of the survey, contextual placement of certain items, and the partial correlation coefficients of the items. These

### *Split Questionnaire Design*

correlations may be readily available in tracking or syndicated studies, because here the researcher knows which (groups of) variables are correlated, from their previous measurements. In cross-sectional studies, prior knowledge about inter-relationships between variables can be obtained from a pilot study. However, even when such prior information is available, the construction of a split questionnaire design such that a minimum amount of information is lost is a challenging task. Since the number of possible split questionnaire designs is exponential in the number of questions, it is not feasible to consider all possible splits in designing a questionnaire for real-life applications. Therefore we suggest, in line with previous practice in marketing research, utilizing the natural structure of the questionnaire, in which questions are placed in blocks. Mostly, several questions measuring, for example, one particular attitudinal or lifestyle trait are administered as a group or block. We use this block-structure to generate split questionnaire designs in two different ways: selecting entire blocks of questions, which we call a “between-block design”, or selecting questions in each block, which we call a “within-block design.” In the between-block design, a “split” is comprised of the allocation of selected blocks of questions and respondents answer all questions in these blocks; in the within-block design, a split is comprised of sets of selected questions in each of the blocks, and respondents answer only those questions in each block. For the first method, given the coherent interpretation of the questions in one block, the problem then simplifies to how these blocks should be administered to respondents in an optimal way. On the other hand, for the within-block design, we need to optimally choose questions in each block. The choice between the within-block and the between-block

design should be based on substantive issues, as well as statistical properties of the two types of designs, as will become clear in the following of chapter. We focus on the problem of how to best develop a split questionnaire and propose a method to optimally choose the splits (a set of blocks of questions or questions in each block offered to a respondent) in this chapter.

### **3.1.2 Outline of the Chapter**

The main contribution of this chapter is to propose a method to design split questionnaires. We apply the modified Federov algorithm to find the optimal design from all possible designs because of its speed and reliability. This method has been previously applied in a different context in the design of conjoint experiments (Kuhfeld, Tobias and Garratt, 1994). We propose using Kullback-Leibler (KL) distance between the complete and split questionnaire data as an optimization criterion. The algorithm searches the candidate splits for the split that is optimal in terms of the given criterion. As explained above, we study both between-block and within-block split questionnaire designs. The split questionnaire, once administered, results in data missing by design, which may result in lack of identification of all parameters from the observed data (Little and Rubin, 1997; Rassler, 2002). Specific overlap of the splits of the questionnaire may help to avoid that identification problem. We explain how to construct identified split questionnaire designs, and how to impute the missing data with the Gibbs sampler. Using a small simulated questionnaire, we enumerate all possible designs and compare that with the result of our design generating

### *Split Questionnaire Design*

algorithm, which reveals that it recovers the optimal split in all cases. We compare the efficiency of split questionnaires generated with our procedure to (random) matrix sampling designs on synthetic data. In practice, market research companies design split questionnaires by randomly choosing blocks, or questions within each block. These methods are similar to the multiple matrix sampling techniques used in testing theory (Shoemaker, 1973), and therefore constitute an appropriate benchmark.

We then apply our approach to data obtained from a questionnaire on web attitudes and perceptions (Novak, Hoffman, and Yung, 2000) to empirically assess the performance of optimal between- and within-block designs, and to compare them to matrix sampling designs and heuristic designs constructed based on a principal components analysis of pilot data. We investigate the sensitivity of the optimal split questionnaire designs to changes in the prior parameters from the pilot study. Finally, we investigate the extent to which the proposed split questionnaire design method may result in better data quality than the complete questionnaire, by studying respondent burden, boredom, and fatigue in a field application of the web-attitude questionnaire. Our conclusion is that optimally splitting questionnaires is worth consideration due to improved questionnaire efficiency and the resulting data quality.

The subsequent sections are organized as follows: Section 2 examines issues in designing a split questionnaire. Since split questionnaire design is one of the methods of collecting data missing by design, we explain other methods of data collection missing by design in Section 3. In Section 4, the design criterion is introduced; the modified Federov algorithm and the

construction of identified split designs are explained. In Section 5, we discuss multiple imputations of the missing data and the estimation of the fraction of missing information. Section 6 provides a simulation study, which investigates the performance of the proposed split questionnaire design method, Section 7 provides the empirical application, and Section 8 summarizes the field study. Finally, in Section 9, the results of this research are discussed and concluding remarks are offered.

### **3.2 Constructing the Split Questionnaires**

Finding an optimal design for a split questionnaire involves finding the configuration of question sets (i.e. those questions given to one respondent, or a “split”) such that a minimum amount of information is lost as compared to the complete questionnaire. The design of a split questionnaire, as we propose it, involves two steps. First, one needs to assign questions to blocks with homogeneous content. Second, one needs to allocate either selected blocks to splits, or selected questions within blocks to splits, resulting in between- and within-block designs, respectively. In the first step, one wants to keep thematically closely related questions in the same block<sup>4</sup>. Raghunathan and Grizzle (1995) call this the contextual placement of questions. We start from the assumption that the questionnaire already consists of a number of blocks with questions that

---

<sup>4</sup> A block structure, if not available a-priori, can be generated using cluster analysis of a pilot with the full questionnaire (Rassler 2002).

### *Split Questionnaire Design*

need to be kept together, and we will utilize that natural structure of the questionnaire. Our approach is thus very suitable for questionnaires comprised of items to measure several multi-item constructs. These are very common in marketing research. Each split questionnaire design is defined by three sets of parameters: the number of splits, the number of blocks/questions per split, and the sampling fraction responding to each split. In this study we investigate the first two parameters and assume throughout that splits are distributed randomly and evenly to respondents. We propose to choose splits from all possible combinations of blocks (between-block designs) or from all possible combinations of questions in each block (within-block designs), using the Kullback-Leibler distance as a measure of information loss, computed from prior parameter estimates. Split questionnaires are one of the methods of collecting data missing by design in surveys with long questionnaires. Now, we explain other methods of data collection that give rise to data missing by design, in order to gain a broader perspective.

### **3.3 Data Missing by Design**

Data collection through surveys requires significant amounts of time, money, and effort. Since time, money and subjects are scarce, in various research areas including marketing, researchers have begun developing more advanced methods to more efficiently collect data. Under time, subject and cost limitations, market research companies sometimes prefer to collect data missing by design, which is also called “planned missingness.” In these studies, companies select sub-parts of the whole

questionnaire to reduce the cost of a study. If planned missingness methods are applied successfully, missingness has little effect on the precision of the parameter estimates of interest. In this section, we talk about these proposed approaches, which are collecting data missing by design. In addition to collecting data missing by design, another currently used procedure in marketing is data fusion, which allows merging data from different sources. Since data fusion and split questionnaires are related, we also discuss data fusion and explain the relationship below. A split questionnaire survey design results in data that is missing by design. Alternative methods are two-stage designs, matrix-sampling designs, subsampling, time-sampling designs, and some experimental design procedures from classical statistics, such as fractional factorial designs or incomplete block designs.

We saw the first applications of data missing by design in experimental psychology and in agricultural experiments (in which plots are used), in which different subsets of questions, plots, or stimuli are administered to different persons, e.g. factorial designs. Since factorial designs take less time (or require fewer resources) and the respondent's task is shorter and less burdening, data collection is more efficient. For instance, Hermkens (1983) uses greco-latin square designs for surveys on equality of income. We also see applications of data missing by design in spatial interpolation problems in environmental science, mining, engineering, geology, soil science and hydrology (Le, Sun, and Zidek, 1997). The most common and widespread usage of data missing by design is in educational testing.

### *Split Questionnaire Design*

Calibration and measurement designs in educational testing are often incomplete designs and used in the framework of item response theory (IRT). The researcher decides to administer only a subset of the total items to the subjects because of the limited testing time (not all available items can be administered to every student). The three commonly used incomplete designs are random incomplete designs, multistage testing designs, and targeted testing designs. In random incomplete designs, the researcher decides which test form is taken by which students without using any a priori knowledge on the ability of a student. In multistage testing designs, the assignment of students to subsets of items from the total item pool in a specific testing stage is based on the observed responses in the previous stage (this is one kind of two-stage design). In targeted testing designs, the structure of the design is determined a priori on the basis of background information. There are two alternative applications of this method. First, the background variables (demographic or income information, etc.) are only used in the assignment of items or tests to students and not in the sampling of the students. In the second application, the background variables are used in the sampling of students as well as in the assignment of tests to students. The efficiency increases if we use a priori knowledge about the difficulty of the items and the ability of students to allocate students to subsets of items (Lord, 1980), which would call for Bayesian methods to develop these kinds of designs. Adaptive or tailored testing in educational testing is another application of “data missing by design.” In adaptive testing, the examinee’s preceding responses are used to select each next item to administer. For example, an examinee answering items correctly would be administered successively more difficult

items, and an examinee answering incorrectly would be administered successively easier items. Although the concept of a “correct” answer may or may not be of use in marketing surveys when designing questionnaires, we believe that the ideas in item response theory models (IRT) from the educational testing literature can be useful in the design of online marketing questionnaires in the future. Some studies on questionnaire designs from item response theory literature are van der Linden, et al. (2004), van der Linden (1999), van der Linden (2004), van der Linden, et al. (1998), and Veldkamp (2002).

The most prominent questionnaire design applications in marketing are conjoint questionnaire designs. Researchers traditionally have constructed designs for (ratings or choice) conjoint experiments using methods from the experimental design literature. Fractional factorial designs are the main method used in experiments. For instance, Lenk et al. (1996) present results that provide shorter questionnaires for metric conjoint analysis. They describe the problems associated with long questionnaires and call for experimental designs and estimation methods to recover parameters with shorter questionnaires. Their paper considers two experimental designs: one in which each subject receives the same set of questions, and one in which subjects receive different blocks of a fractional factorial design. Based on research by Huber and Zwerina (1996), Sandor and Wedel (2001) design conjoint choice experiments based on prior information about the parameters and their associated uncertainty, elicited from managers. They use Bayesian design procedures that assume a prior distribution of

### *Split Questionnaire Design*

likely parameter values and optimize the design over that distribution. Apart from conjoint questionnaire designs, there is to date no research on collecting data missing by design in surveys, and we intend to fill this void with this chapter. Before explaining some alternative tools for collecting data missing by design, we provide some differences and similarities between questionnaire design for conjoint experiments and survey designs.

Conjoint experiments (conjoint questionnaire design) are a specific instance of experimental design, whereas split questionnaire designs toll within the value of survey designs. An experimental design specifies how to allocate resources (attribute levels in conjoint experiments that we want to learn consumer's preferences) in the study. On the other hand, sampling is an economical way to select a small part of the population, so that study of that part permits broad generalizations within reasonable limits of doubt. In this chapter on split questionnaire designs, our purpose is to generate different versions of the questionnaire (which contain fewer questions than the complete questionnaire) with minimum information loss and we would like to know which questions from the whole questionnaire should be chosen to be administered together. The main difference, compared to survey sampling, is that instead of sampling subjects, we select questions and distribute them evenly to subjects. Sufficient subjects should respond to these different versions of the questionnaire. Our approach for designing split questionnaire designs can be modified by selecting questions based on sampled subjects' background information or depending on some selective (classifying) questions. The issue of how many subjects should respond to each version of the questionnaire is also an important issue for future research. After we generate optimal split questionnaires, we collect

data and impute the missing parts. There is no imputation in conjoint designs after data collection. To design conjoint questionnaires, an important assumption that is often made is to assume zero values for the attribute weights. However, in designing split questionnaires, we use the covariance of the questions obtained from pilot studies.

### **3.3.1 Two-Stage Designs**

Two-stage designs are the most common example of procedures that generate data missing by design. The first stage consists of core questions to elicit information which we want to have from all respondents, whereas the second stage, the remaining blocks of questions, are given to a subset of the entire sample, or to a stratified random sample with selection probabilities dependent on the first stage. The correlation between the core measure in the two stages and selection criteria for the second stage sample provide the information needed to make full-sample inferences about the second stage measures (Neff, 1996).

### **3.3.2 Matrix Sampling Design**

Matrix sampling refers to the random sampling of a rectangular array of row-column entries from a larger matrix from the population. The National Assessment of Educational Progress (NAEP) uses matrix sampling designs in item testing<sup>5</sup>. Item testing is a popular psychometric application of this

---

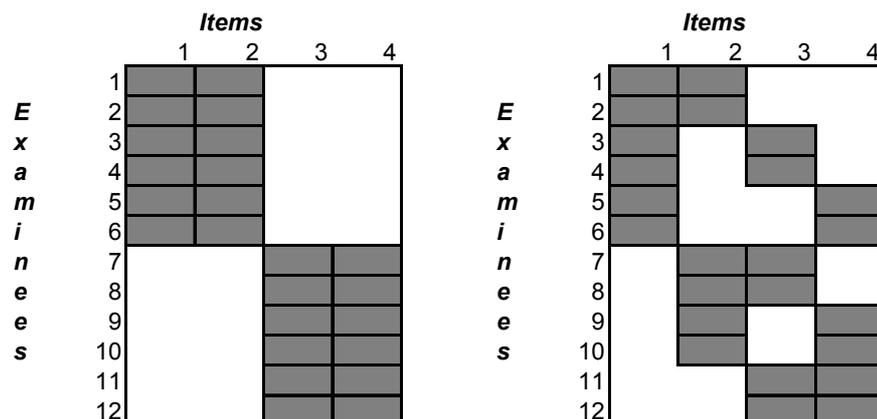
<sup>5</sup> NAEP is the only national assessment in the US that measures what American students know and can do over time in various subjects such as reading and mathematics. The analysis of NAEP is IRT based and contains several consecutive steps.

### *Split Questionnaire Design*

method in which the rows constitute examinees and the columns constitute items. If more than one matrix is sampled, this is referred to in the literature as a multiple matrix sampling. Figure 3.1 depicts a non-overlapping multiple matrix sampling design (NMS) wherein examinees and items are sampled without replacement, and an overlapping multiple matrix sampling design, which in fact is a balanced incomplete block design (BIB). When more than one matrix is sampled, the point estimates for a single matrix are repeated, computed and averaged over all matrices, since the mean of unbiased estimates is also unbiased. When it comes to the computation of standard errors, the situation is more complicated. In matrix sampling designs, different respondents are asked different questions, and the set of questions varies across strata. These designs have common applications in computer-aided interviewing, or as a part of item experiments. In split questionnaire survey design, missing data are imputed to end up with a complete data set, which is not the case in matrix sampling as used in educational testing.

Figure 3.1: An example of NMS and BIB design

1-) Non-overlapping matrix sampling (NMS) 2-) Balanced incomplete block design (BIB)



### 3.3.3 Time Sampling Design

There are two basic ways to obtain information about the continuous behavior of consumers in time. The first is continuous consumer panels, which help to obtain a continuous record of the behavior of consumers for the entire time period. The second is to sample time, that is, to observe consumers at various points in time and to infer from these observations what behavior took place for those periods for which no measurements were made. Among market researchers, the most commonly used method is to have each sampling unit record its own continuous behavior via a self-administered form, usually referred to as a diary. Although used in a wide variety of contexts, the most frequently used types of consumer panels in

### *Split Questionnaire Design*

marketing are the purchase panel, the media measurement panel and the product test panel.

Sampling over time enables us to monitor, analyze and understand social processes through the estimation and analysis of changes in variables of interest. In addition to the usual sample design issues considered for a sample used for one time period, the design of a time sampling scheme needs to consider the frequency of sampling and the spread and pattern of inclusion of selected units over time. A key issue is whether to use overlapping or non-overlapping samples over time. For overlapping samples, the precise pattern of overlap must be designed. Factors that affect the design of a sample over time are: the key estimates to be produced, the type and level of analyses to be carried out, cost, data quality, and reporting load. The interaction between the design of the sample in time and the other features of the design, such as stratification and cluster sampling, also needs to be decided. Time series may be produced and analyzed, which may involve seasonal adjustment and trend estimation. Composite estimation is one of the methods of estimation that is used in time sampling that involve using data for the current and previous time periods and give different weights to matching and non-matching sample units.

Repeated, panel, and longitudinal surveys, rotating panel surveys, split panel surveys and rolling samples are important examples of the application of time sampling. A longitudinal survey (or panel survey) is a survey that uses a sample in which the same units are included for several time periods. A repeated survey is a survey conducted at different times

with no attempt to have sample units in common. Rotating panel survey is a panel survey in which a proportion of units are removed from the survey at some time periods and replaced by other units. In this method, a different rotation pattern can be used, i.e. the pattern of inclusion of sample units over time, such as overlapped or nonoverlapped (orthogonal) patterns. One example of time sampling design is illustrated in Figure 3.2.

### **3.3.4 Subsampling or Multistage Sampling**

In subsampling, the aim is to divide the blocks into smaller and preferable subsamples (Figure 3.2). If the blocks represent clusters, subsampling is generally used to divide larger clusters into smaller clusters in sampling design. The advantage of this method is decreasing variance due to a decrease in the degree of clustering, without incurring a proportional increase in cost (Kish, 1965). The difference between estimates from these independent samples may be used to estimate the error variance. Then these error variances are straightforwardly projected to the entire sample formed from the combined subsamples. Subsampling designs require a minimum of two subsamples and homogeneity between samples. Although it is often not practical to include many more than two in the design, this method can be extended to more stages.

### **3.3.5 Data Fusion**

Data fusion is related to split questionnaire designs. Data fusion or statistical file matching techniques merge data sets from different survey

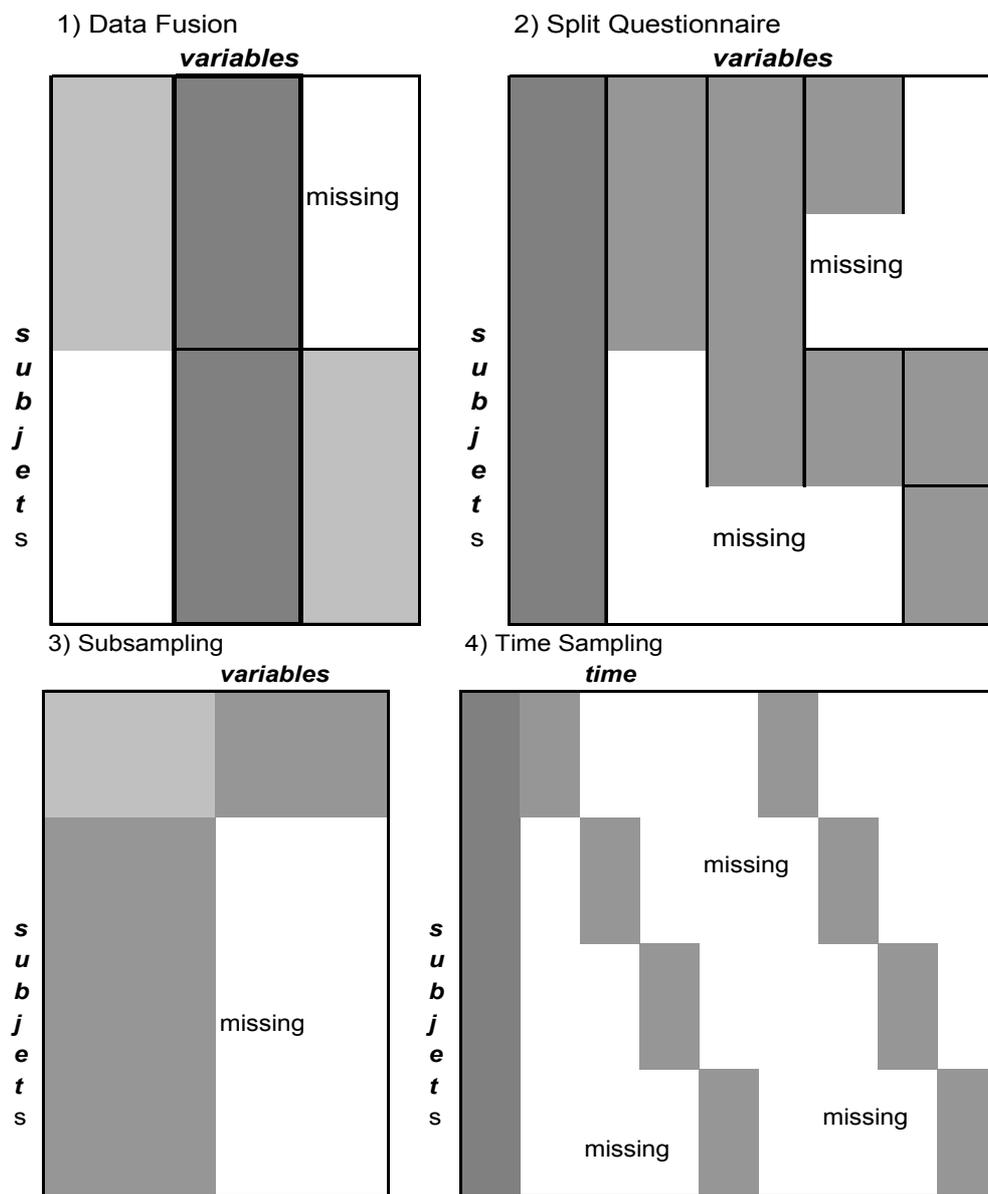
### *Split Questionnaire Design*

samples to solve the problem that exists when no single file contains all variables of interest (Figure 3.2). Split data arise when data on two different sets of variables are obtained from two independent samples, while a number of variables (usually demographics) are measured in both samples (Kamakura and Wedel, 1997, 2000, and Gilula et al., 2006). Merging data sets is usually done on the basis of variables common to all files, and the methods in question assume conditional independence of the variables never jointly observed given the common variables.

Data fusion can also be used to reduce the required number of respondents or questions in surveys. For example, the Belgium National readership survey on media and products is distributed to two different groups of 10,000 respondents each in Belgium and later merged into a single survey (van der Puttan et al., 2002). In this way, the cost and the time for each respondent to complete the questionnaire was conveniently reduced.

The focus in data fusion studies, however, is more on how to merge two different data sets from different surveys. But in principle, we can design split questionnaires and distribute them to respondents and later merge these different sets using data fusion techniques. Split questionnaire survey design can be applied especially for media and purchase surveys. For instance, data from a television measurement panel and a purchasing behavior panel can be merged together with data fusion.

Figure 3.2 Data missing by design



### **3.3.6 Incomplete Block Design**

The goal in incomplete block design is to construct a design such that any pair of treatments (blocks of questions) occurs equally often within some block (split). A solution can be found for any number of treatments and any size of block, but most of the solutions require too many replications for the usual situations in survey designs. For a given number of treatments and a given size of incomplete block, balanced designs allow little choice in the number of replications. Assignment of blocks of questions to splits can depend on the following constraints: The number of blocks assigned to each split ( $k$ ), the number of splits to which each block assigned ( $\lambda$ ), combinations of blocks are assigned to splits (a minimum number of splits or number of split per pair, etc.).

When an incomplete design is formed so that every pair of treatments occurs together in the same number of blocks, the design is called a balanced incomplete block design (Giesbrecht, 2004). A balanced incomplete block design (BIBD) is expressed with five parameters, ( $v$ ,  $b$ ,  $r$ ,  $k$ ,  $\lambda$ ), and is a family of  $b$  sets, called blocks, each consisting of  $k$  (where  $k < v$ ) elements taken from a set of  $v$  elements, such that each element occurs in exactly  $r$  blocks and every pair of elements occurs together in exactly  $\lambda$  blocks. Since  $b$  and  $r$  can be calculated from  $v$ ,  $k$  and  $\lambda$ , we use ( $v$ ,  $k$ ,  $\lambda$ ) as the parameters for the design. Balanced incomplete block (BIB) designs do not exist for all combinations of blocks sizes ( $k$ ), numbers of treatments ( $v$ ) and number of replications ( $r$ ). There are four necessary conditions for the existence of BIB design.

1.  $vr = bk$
2.  $\lambda(v - 1) = r(k - 1)$
3.  $b \geq v$
4. if  $v = b$ , and

if  $v$  is even then  $k - \lambda$  is a perfect square and if  $v$  is odd then  $z^2 = (k - \lambda)x^2 + (-1)^{(v-1)/2} \lambda y^2$  has a solution in integers with  $x, y, z$  not all equal to 0.

Proof: The number of pairs in a block is  ${}_k C_2 = \frac{k(k-1)}{2}$ . The number of treatment pairs is  ${}_v C_2 = \frac{v(v-1)}{2}$ . There are  $b$  blocks, so the total number of pairs is  $b \times \frac{k(k-1)}{2}$ . Each pair occurs  $\lambda$  times, so the total number of pairs is  $\lambda \times \frac{v(v-1)}{2}$ . Equating these two expressions gives  $r.(k - 1) = \lambda.(v - 1)$

A BIB with 20 blocks would lead to 190 version of the questionnaire for identifiability, and would necessitate unrealistically large sample sizes. If instead a partial BIB is utilized, this leads to many occurrences where questions/blocks do not co-occur, hence, bivariate information is not always available (which leads to identification problems).

We use prior information to design split questionnaires, but incomplete block design is not based on such prior information. Balanced incomplete

*Split Questionnaire Design*

block designs depend on the number of blocks assigned to each split ( $k$ ) and the number of splits to which each block assigned ( $\lambda$ ). We use covariance relationships as a prior to generate split questionnaire designs, which allow us to reduce more questions relative to BIB designs. Incomplete balanced block design comes with some certain number of replications (distinct splits) and need more splits (different versions of questionnaires) for identification. Because in split questionnaire designs, we don't have any restrictions on the number of splits in generating identified designs, we need fewer splits.

Figure 3.3: Feasible balanced incomplete block design

Splits	Blocks				
	1	2	3	4	5
1	1	1	0	0	0
2	1	0	1	0	0
3	1	0	0	1	0
4	1	0	0	0	1
5	0	1	1	0	0
6	0	1	0	1	0
7	0	1	0	0	1
8	0	0	1	1	0
9	0	0	1	0	1
10	0	0	0	1	1

One of the applications of incomplete balanced block design in marketing is demonstrated by Rink (1987). He explains and illustrates how these designs can circumvent problems where the respondent must rank many objects. Raghavarao and Federer (1979) present balanced incomplete block design designs as an alternative approach to the randomized

response method for dealing with sensitive questions in a survey context. Their proposed method increases the chances of obtaining honest and unbiased responses by protecting respondents' privacy in a survey, which includes questions that the respondent may not be inclined to answer truthfully. Each respondent is administered a questionnaire containing a subset of the possible questions in these designs. That is, each respondent is assigned a "block" in an incomplete block design<sup>6</sup>. This method applies to questionnaires in which all blocks have at least one quantitative question. The key idea in this approach is that scores for a set of  $k$  of the  $v$  questions, sensitive and/or non-sensitive, are added, and only a total score for the  $k$  questions is reported by the respondent. Different respondents receive different sets of  $k$  questions; there are  $b$  different sets of questions constructed according to known experimental designs, such as the supplemented block designs and balanced incomplete block design. The block of  $k$  questions is randomly assigned a respondent, and all blocks have an equal or nearly equal number of respondents. We can estimate population proportions or means for each question from the block totals; however, we are unable to determine what an individual's response was to a particular question. With the usage of incomplete balanced block designs, one saves interviewing time for questionnaires with several sensitive questions and potentially improves response.

---

<sup>6</sup> In a split questionnaire design, each split (i.e. version of the questionnaire) plays the role of a "block."

## **3.4 Measuring Information Loss**

### **3.4.1 Optimal Split Questionnaires Using KLD**

We use the Kullback-Leibler (KL) measure, the distance between two probability models, to choose the best among all possible designs. The KL-distance was developed by Kullback and Leibler (1951) from “information theory.” Here, it is first applied to design construction. The KL-distance defines the distance between the probabilistic models  $f$  and  $g$  for as the (usually multi-dimensional) integral:

$$l(f,g) = \int f(y) \log\left(\frac{f(y)}{g(y|\theta)}\right) dy \quad (3.1)$$

$l(f,g)$  is the “information” lost when  $g$  is used to approximate  $f$ . An equivalent interpretation of minimizing  $l(f,g)$  is finding an approximating model that is the shortest distance away from “the truth.” If  $f(y)$  and  $g(y|\theta)$  are multivariate normal distributions with a common variance-covariance matrix, then the Kullback-Leibler distance reduces to the Mahalanobis distance (Bar-Hen and Daudin, 1995), which is frequently used as a distance measure in the literature.

We assume that the optimization of the split questionnaire design (SQD) is done under one external constraint fixed by the researcher, which is the total number of splits ( $K$ ) desired. We assume that the researcher knows this number from prior considerations, or that issues related to the implementation of the questionnaire dictate it. The optimization can also accommodate any other practical constraint, such as one that induces

respondents to answer a fixed number of (blocks of) questions, i.e. each candidate split should contain a predetermined number of blocks. These constraints are illustrated below. After generating  $K$  splits and evenly distributing these splits to respondents, the Kullback-Leibler distance is calculated. In our notation,  $K$  denotes the total number of splits,  $N$  is the number of respondents,  $B$  is the number of blocks,  $Q_b$  is the number of questions in block  $b$ ,  $Q$  is the total number of questions, ( $\sum_{b=1}^B Q_b = Q$ ),  $Y$  is the data-matrix containing the answers of the respondents and  $D$  is the questionnaire design matrix with 0/1 entries (i.e. a fully observed matrix of indicators whose elements are zero or one depending on whether the corresponding elements of  $Y$  are missing or observed):

$$d_{ij} = \begin{cases} 1 & \text{if question } j \text{ is given to respondent } i \\ 0 & \text{otherwise} \end{cases}$$

Now  $f(Y|D)$  is the likelihood of the incomplete data with respect to the split questionnaire design matrix and  $f(Y)$  is likelihood of the data with respect to the complete questionnaire. The Kullback-Leibler distance between the complete data likelihood  $f(Y)$  and the split data likelihood  $f(Y|D)$  is defined as:

$$\begin{aligned} \text{KL}(D) &= \int f(Y) \ln \left[ \frac{f(Y)}{f(Y|D)} \right] dY, \\ &= E \ln[f(Y)] - E \ln[f(Y|D)], \end{aligned} \tag{3.2}$$

### *Split Questionnaire Design*

where each expectation is with respect to the true distribution  $f(Y)$ , where  $Y_{N \times Q} = [Y_1, Y_2, \dots, Y_Q]$ . Thus, the  $KL(D)$  in this case measures the distance between the distribution of the complete data  $f(Y)$  and the incomplete data  $f(Y|D)$  given the split questionnaire design  $D$ , i.e. it assesses the expected loss of information by deleting data according to the split questionnaire, relative to the complete questionnaire data. The most efficient questionnaire design ( $D$ ) minimizes  $KL(D)$ . The first term on the right hand side in the equation for  $KL(D)$  is the same for each possible design since it is derived from the complete questionnaire. Consequently, maximization of the second term on the right hand side suffices. Since  $f(Y)$  is the same for each possible design,  $\text{Inf}(Y|D)$  will be maximized in the sequel. Minimizing the KL-distance can be seen as finding the split questionnaire yielding incomplete data, which are closest in expectation to the data that would have been obtained with the complete questionnaire.

We will assume the form of  $\text{Inf}(Y|D)$  to be a multivariate normal, as a function of the parameters  $\mu$  and  $\Sigma$ , as shown below. In Appendix I we provide an extension of the KL-distance for mixed data consisting of continuous and discrete variables using a general location model. However, multivariate normality is often assumed for responses of scales in many marketing surveys, including those measuring attitudes, satisfaction, lifestyles etc. (Huber et al. 1993). In addition, the normal distribution has minimal KL-distance to any unknown distribution function (O' Hagan 1994), and in this case minimizing the KL-distance is equivalent to minimizing the Mahalanobis distance.

We have  $Q$ -variate normal data  $N_Q(\mu, \Sigma)$  with  $\mu = (\mu_1, \dots, \mu_Q)$  and  $\Sigma_{Q \times Q}$ . For now,  $\mu_{Q \times 1}$  and  $\Sigma_{Q \times Q}$  are assumed known. These are considered prior information that can be obtained from past data or through a pilot experiment. The aim is to construct the design using  $\mu_{Q \times 1}$  and  $\Sigma_{Q \times Q}$  as prior information. Thus, we have the following optimal design criterion:

$$L = \ln L(Y | D, \mu, \Sigma)$$

$$= \prod_{i=1}^n \left( \frac{-p_D}{2} \right) \ln(2\pi) - \frac{\ln |\Sigma(D)|}{2} - \frac{1}{2} [(Y_{\text{obs}} - \mu(D))' \Sigma(D)^{-1} (Y_{\text{obs}} - \mu(D))] \quad (3.3)$$

where  $p_D$  is the number of parameters under design  $D$ ,  $n$  the total number of respondents,  $Y_{\text{obs}} = Y_{ij} d_{ij}$  the data observed under the split questionnaire  $D$ , and  $\mu(D)$  and  $\Sigma(D)$  denote the subvector of the mean vector  $\mu$  and the square submatrix of the covariance matrix  $\Sigma$  which are obtained from complete data estimates from a pilot study, respectively, that pertain to the variables that are observed in design  $D$ .

### **3.5 Identification Issues in Constructing SQD**

When we construct a split questionnaire design, we should be able to estimate all parameters from the observed incomplete data. We call a design that enables the estimation of all parameters (of the multivariate normal distribution) a fully identified design. Clearly, not all designs are fully identified. We illustrate the identification problem briefly through the following example. Assume we want to estimate the parameters of a

### *Split Questionnaire Design*

multivariate Normal distribution for three blocks, X, Y and Z in a between-block design. However, we have a split A- with only X and Y and a split B- with only X and Z observed together. The covariance matrix of Y and Z is written  $V(Y,Z) = V(Y,Z|X) + V'(X,Y)V(X)^{-1}V(X,Z)$ , where  $V(X)$  the covariance matrix of X, and  $V(Y,Z|X)$  the covariance matrix of Y and Z conditional on X. We can estimate  $V(X,Y)$  from split A,  $V(X,Z)$  from split B, and  $V(X)$  from both splits, but we cannot only directly estimate  $V(Y,Z|X)$  from the available incomplete data. However, if we assume conditional independence of the Y and Z variables given X, we can estimate  $V(Y,Z)$  from  $V(Y,Z) = V'(X,Y)V(X)^{-1}V(X,Z)$ , since all terms on the right hand side are estimable (see Gilula, McCulloch and Rossi, 2006; Rassler, 2002; Rodgers, 1984). However, if we use this conditional independence assumption in a model for imputing the missing data, this implies that for all parameter estimates or statistics subsequently computed from the imputed data this conditional independence assumption should also hold. That assumption is a strong one, which may limit the usefulness of such split questionnaire designs in practice.

Rassler (2002) and Gilula, McCulloch and Rossi (2006) suggest (in the context of data-fusion) to use informative priors in the imputation to overcome the identification problem. The use of priors adds information that enables estimation of the parameters that are not identified by the split questionnaire design. The fact that  $V(Y,Z|X)$  is inestimable results in non-positive definite variance-covariance matrix  $V(X,Y,Z)$ , which we can avoid using prior information. If one uses the Gibbs sampler for imputation, as we will below, such prior information also overcomes lack of convergence.

Using informative priors for the means and covariance matrix of the normal distribution results in an imputed dataset devoid of conditional independence properties induced by the design, which is highly desirable. Since the design itself is constructed based on such prior information, it is natural to also include that same prior information in imputing the missing data. However, it is even more desirable to address the identification problem by constructing designs that do not suffer from it, which we do below.

If all possible pairs of questions occur in an optimal split questionnaire design, this ensures that all parameters of a multivariate normal distribution are identified and estimable from the observed data. Let us consider the between-block design: if we have a questionnaire with  $n_B$  blocks and we impose the constraint of  $n_S$  blocks per split, then the number of splits  $K$  for a fully identified design needs to satisfy  $\binom{n_B}{n_S} \leq K \leq \frac{n_B(n_B - 1)}{n_S(n_S - 1)}$ , where  $\binom{n_B}{n_S}$  is the size of the candidate split-set. Note that this is a necessary, but not sufficient condition. In practice one can easily check the identification of any design by looking at the  $(D'D)$  matrix: only designs with all off-diagonal elements greater than 0 are fully identified designs. In generating constrained split questionnaire designs, we recommend that one only considers fully identified designs by imposing the identification constraint  $(d_i' d_j) \neq 0, \forall_{i \neq j}$ , and employ the prior information used to construct the design also in imputing the missing values. This is what we will do

throughout the remainder of this chapter, and we recommend it in general as a procedure for constructing split questionnaires.

### **3.6 Design Generating Algorithm**

We assume that the split questionnaire design (SQD) is constructed under the external constraint that the total number of splits ( $K$ ) is fixed. The optimization can also accommodate other practical constraints, such as that one or more blocks are included in every split, or that each candidate split contains a predetermined number of blocks. Note that these constraints are possible, but not needed (such constraints are illustrated in the applications below). In order to find the most efficient  $K$  splits out of all possible candidate splits ( $N_S = 2^Q$ , with  $Q$  the number of questions), one could generate all  $N_D = \binom{2^Q}{K}$  possible designs and retain the one with the smallest value of the KL-measure. In most practical situations, it is not feasible to do this, since it is usually not computationally feasible to list all  $N_D$  possible subsets out of  $2^Q$  designs. Therefore, we need to use an efficient algorithm to search the design-space. Such an algorithm would conduct a search among all possible candidate splits for one that improves the KL criterion. We apply the modified Federov algorithm for that purpose. The modified Federov algorithm is a popular algorithm for experimental design construction, since it is robust and fast. Kuhfeld, Tobias and Garratt (1994) applied it to generate conjoint choice designs.

We start describing the procedure that is used to generate the between-block designs. We assume that if there are  $N$  individuals, then  $N/K$

individuals will be assigned randomly to each of the  $K$  splits. Each alternative split questionnaire design then consists of an  $N \times Q$  matrix  $D$  with  $K$  different split patterns. Each entry in the matrix  $D$  is a 0 or 1, indicating whether a question is included or excluded in that particular split. In constructing between-block designs, we constrain all questions in one block to be assigned to the same respondent. That is, if we have five blocks with four questions and one particular split at the block-level is [11010], we will use  $d_{ij}=[1111 \ 1111 \ 0000 \ 1111 \ 0000]$  as a row in the design matrix  $D$ . The proposed procedure to construct split questionnaire designs operates as follows:

**Step 1.** Build a candidate split-set ( $C$ , a  $N_S \times Q$  matrix), which is a list of all potential splits contained in its rows. Inadmissible designs are removed from  $C$ .

**Step 2.** Choose a starting design at random, say  $D_0$ . Using the pilot data, obtain estimates for the parameters of the model for each of the questions in the questionnaire. Compute the KL-measure for the starting design  $KL(D_0)$  based on these estimates, using (3.3).

**Step 3.** Take the first split (first  $N/K$  rows) in the starting design  $D_0$ . Exchange that with the candidates,  $\ell = 1, \dots, N_S$ , i.e. each of the rows in  $C$ , in turn. For every exchange, compute the KL-distance in (3.3), i.e.  $D_0^\ell$ . Keep that split that minimizes the KL-distance, i.e.  $D_1 = \min_{\ell} KL(D_0^\ell)$ , and replace  $D_0$  by  $D_1$ .

### *Split Questionnaire Design*

**Step 4.** Find the best exchange (if one exists) for the next split in the target design  $D_1$  (i.e. the second set of  $N/K$  rows), by sequentially processing the candidates  $\ell = 1, \dots, N_S$  in  $C$ , and replacing the design matrix  $D_1$  by  $D_2 = \min_{\ell} \text{KL}(D_1^{\ell})$ .

**Step 5.** Ensure that the design is fully identified by checking off-diagonals of the  $(D'D)$  matrix at every step, and reject splits that cause zero off-diagonal values.

**Step 6.** The first iteration is completed once the algorithm has found the best exchanges for all of the splits in the target design matrix. Then, the algorithm moves back to the first split in the target design matrix and replaces it again with each candidate in  $C$ , cycling through steps 3 and 5, until no improvement is possible.

**Step 7.** To avoid local optima, the whole process is restarted with different (random) starting designs and the best design is selected, i.e. the one that yields the lowest KL-distance.

#### **3.6.1 Generating Within-Block Designs**

Whereas the construction of between-block designs is feasible with the modified Fedorov algorithm described above, that of the within-block design is not in most practical situations, because of the enormous size of the design space. Therefore, we choose questions within each block using a “greedy” approach, as follows. Instead of optimizing the full within-block split design, we generate splits for each block sequentially. For block  $B$  there are  $2^{Q_B}$  possible splits, with  $Q_B$  the number of questions. We have a

candidate split-set for each block, denoted as  $C_b$ , for  $b=1,\dots,B$ . The procedure then operates as follows.

**Step 1.** Build a candidate split-set ( $C_b$ ), for each block. Inadmissible designs are removed from  $C$ . Choose a starting design at random for every block, say  $D_{0,b}$ .

**Step 2.** Find the optimal  $K$  splits in the first block from  $C_1$  using the modified Federov algorithm as described in the Steps 3-6 above, assuming the other blocks are complete, to obtain  $D_{1,1}$ .

**Step 3.** Then, find the optimal splits in the second block searching across the candidate splits in  $C_2$ , as described in steps 3-6 above, given the optimal splits of the first block and assuming the remaining blocks are complete, to obtain  $D_{2,1}|D_{1,1}$ .

**Step 4.** Continue this procedure by sequentially passing through the remaining blocks, finding the optimal splits for each block using steps 3-6 above, given the optimal designs of the previous blocks, and assuming the remaining blocks complete, thus obtaining  $D_{b,1}|D_{b-1,1},\dots,D_{1,1}$ .

Unfortunately, it proves difficult to produce fully identified within-block designs using the “greedy” approach described. We therefore choose to generate only locally identified designs by checking the  $D_b'D_b$  matrix of each block  $b$  separately. This does not guarantee the appearance of all question-pairs in the complete design, which is needed for the design to be fully identified. Thus, the constructed within-block split questionnaire designs are neither fully identified nor globally optimal, but are still more

### *Split Questionnaire Design*

efficient than designs constructed by choosing questions within each block at random or with heuristic procedures.

For within-block designs, constraints can be imposed by only considering admissible designs in the candidate split set  $C_b$ . One important class of constraints is imposed by forced within-block skip patterns in the questionnaire (see Sudman and Bradburn, 1989, p.224). The within-block branching structure of the questionnaire can be accommodated in the split questionnaire design, by forcing a higher node question into any split that also contains the lower node question.

## **3.7 Multiple Imputations with Gibbs Sampling**

The within- and between-block split questionnaire designs produce datasets with intentionally missing data. To obtain complete data, instead of using a single imputation, which ignores uncertainty due to imputation and therefore underestimates the variability of the resulting estimates (Rubin, 1987), we use Bayesian proper multiple imputations by drawing values of missing data ( $Y_{\text{mis}}$ ), and  $\mu$  and  $\Sigma$  from their full conditional posterior distributions using Gibbs sampling (Gelfand and Smith, 1990). We use informative priors,  $\mu_{\text{pr}}$  and  $\Sigma_{\text{pr}}$ , obtained from the full questionnaire in a pilot study, with  $n_0$  and  $\rho$  the prior number of observations and degrees of freedom on which the  $\mu_{\text{pr}}$  and  $\Sigma_{\text{pr}}$  are based, respectively. Let  $\Sigma_{\text{obs,obs}}$ ,  $\Sigma_{\text{mis,mis}}$ , and  $\Sigma_{\text{mis,obs}}$  denote the sub-matrices of  $\Sigma$  formed by the indices corresponding to the observed and missing  $Y$  values;  $\mu_{\text{obs}}$ ,  $\mu_{\text{mis}}$  denote the corresponding sub-vectors of  $\mu$ . The conditional distribution of  $Y_{\text{mis}}$ , given

$Y_{obs}$ ,  $\mu_m$ , and  $\Sigma$ , is normal distribution with mean  $\mu_{mis} + \Sigma_{obs,mis} \Sigma_{obs,obs}^{-1} (Y_{obs} - \mu_{obs})$  and variance  $\Sigma_{mis,mis} - \Sigma_{obs,mis} \Sigma_{obs,obs}^{-1} \Sigma_{mis,obs}$ . The Gibbs sampler iterates between:

**Step 1.** Draw  $Y_{mis}^{(t+1)}$  given  $\mu_0$ ,  $\Sigma_0$ , and  $Y_{obs}$ :

$$Y_{mis}^{(t+1)} | Y_{obs} \sim MVN\left(\mu_{mis} + \Sigma_{obs,mis} \Sigma_{obs,obs}^{-1} (Y_{obs} - \mu_{obs}); \Sigma_{mis,mis} - \Sigma_{obs,mis} \Sigma_{obs,obs}^{-1} \Sigma_{mis,obs}\right) \quad (3.4)$$

**Step 2.** Draw  $\Sigma^{(t+1)}$  given  $\mu^{(t)}$  and  $Y^{(t+1)} = (Y_{obs}, Y_{mis}^{(t+1)})$  from<sup>7</sup>:

$$\Sigma^{(t+1)} | Y \sim IW(n_{obs} + \rho, (n_{obs} - 1)S + \rho \times \Sigma_{pr} + S_m) \quad (3.5)$$

where S is the sample covariance matrix and

$$S_m = \frac{n_{obs} \times n_0}{(n_{obs} + n_0)} (\bar{y} - \mu_{pr})(\bar{y} - \mu_{pr})'$$

**Step 3.** Draw  $\mu^{(t+1)}$  given  $\Sigma^{(t+1)}$  and  $Y^{(t+1)} = (Y_{obs}, Y_{mis}^{(t+1)})$  from<sup>8</sup>:

$$\mu^{(t+1)} | (\Sigma^{(t+1)}, Y) \sim N\left(\frac{1}{n_{obs} + n_0} (n_{obs} \bar{y} + n_0 \mu_{pr}), \frac{1}{n_{obs} + n_0} \Sigma^{(t+1)}\right) \quad (3.6)$$

---

<sup>7</sup> With noninformative priors:  $\Sigma^{(t+1)} | Y \sim IW(n_{obs} - 1, (n_{obs} - 1)S)$

<sup>8</sup> With noninformative priors:  $\mu^{(t+1)} | (\Sigma^{(t+1)}, Y) \sim N\left(\bar{y}, \frac{1}{n_{obs}} \Sigma^{(t+1)}\right)$

The Gibbs sampler is easy to implement and enables quick imputation of the missing values. In addition, it can be used simultaneously and in the same manner to impute missing values arising to item non-response (Schaffer, 1997).

### **3.8 Estimation of the Fraction of Missing Information**

The incomplete data generated through the split questionnaire design contain less information on the parameters than the complete data. We estimate the fraction of missing information of the parameters using the missing information principle (Orchard and Woodbury, 1972, see appendix B). Since the complete data information is the sum of the observed data information and the missing data information, we can write:

$$\frac{1}{V(\hat{\theta})} = \frac{1}{V(\hat{\theta}_{obs})} + \left( \frac{1}{V(\hat{\theta})} - \frac{1}{V(\hat{\theta}_{obs})} \right) \quad (3.7)$$

Here  $V(\hat{\theta})$  is the complete information on  $\theta$  estimated from the Fisher information matrix.  $V(\hat{\theta}_{obs})$  is the expected observed data information, which we estimate after the multiple imputation of the missing data with the Gibbs sampler. If we divide both sides by the missing information and take the fraction of missing information ( $\gamma$ ) to be equal to the missing information divided by the complete information, we obtain:

$$\gamma = \frac{\left( \frac{1}{V(\hat{\theta})} - \frac{1}{V(\hat{\theta}_{\text{obs}})} \right)}{\frac{1}{V(\hat{\theta})}} \quad (3.8)$$

This quantity shows how much information there is in the data on the parameters in question, and can be used as a statistic to evaluate the efficiency of split questionnaire designs.

### **3.9 Simulation Studies**

Before we extensively investigate the performance of split questionnaire designs on empirical data below, we first illustrate them with simulated data. We conduct two simulation studies, focusing on between-block designs. First, we investigate the performance of the modified Fedorov algorithm in identifying the optimal design. Second, we compare optimal split questionnaire designs to matrix sampling designs.

We construct a split questionnaire design that is small enough to enumerate all possible designs, which makes it possible to investigate the performance of the modified Fedorov algorithm in finding the optimal design. Let  $Y_{ij}$  denote the answer of respondent  $i \in \{1, \dots, N\}$  to question  $j \in \{1, \dots, Q\}$ , which forms the complete data matrix  $Y$ . We assume a between-block design, with  $B = 5$  blocks and each block containing  $Q_b = 4$  questions, so that in total we have twenty questions. We generate  $Y$  from a multivariate normal distribution with given  $\mu_{Q \times 1}$  and  $\Sigma_{Q \times Q}$ . The matrix  $X$  is an

### *Split Questionnaire Design*

$N_S \times B$  matrix containing  $N_S$  possible or candidate splits, 1 denoting an included block and 0 denoting an excluded block. There are 32 candidate split points contained in the matrix  $X$ , but unrealistic or undesirable combinations such as one where none of the questions is asked (a row with only zeros in the design matrix  $X$ ) or where just one block of questions is asked, are excluded, as indicated in the candidate split set shown in Table 3.1. Even under the external constraint that fixes the number of desired splits ( $K$ ), there are many possible designs. For example, there are in total 5311735 ( $= 26!/(16!10!)$ ) different designs for  $K = 10$  splits. We choose  $K$  splits from the candidate split matrix in Table 3.1, and distribute these splits evenly to one hundred subjects. We do this both with the modified Federov algorithm and through complete enumeration. The matrix  $D$  contains the design with the  $K$  splits. We eliminate the responses of the subjects from the complete data matrix ( $Y$ ) according to the split design ( $D$ ) and compute the KL distance. We choose the SQD design with the maximum  $\text{Inf}(Y|D)$  among all possible designs as the optimal design. We investigate three different numbers of desired splits:  $K = 5$ ,  $K = 10$  and  $K = 15$ .

The time that the modified Federov algorithm needed to find the optimal questionnaire design with  $K=5$ , 10 or 15 splits is compared to that for complete enumeration in Table 3.2. All calculations are done with a Pentium 3 computer, using GAUSS software. For the Federov algorithm, we used 10 iterations, and 1000 different random starts. All 1000 random starts produced the same optimal design in all three cases in  $1/10^{\text{th}}$  or less of the computation time of complete enumeration, as shown in Table 3.2. This indicates that the performance of the Federov Algorithm as applied to the problem of split questionnaire design is highly satisfactory.

We now illustrate the performance of optimal between-block split questionnaire designs (SQD) relative to matrix sampling designs (MSD) in a second simulation study (within-block designs are investigated more extensively in the empirical application below). We have six blocks and five questions per block. We optimally design the questionnaire and impute the resulting missing data with the Gibbs sampler. We investigate constrained and unconstrained between-block designs, with 5 or 10 splits. To assess the performance of the proposed method, next to the fraction of missing information, we compute the KL-distance and the Bayes information criterion (BIC), where  $BIC = -2 \times \ln f(Y|D) + \ln(N) \times 2$ . Further, we calculate the mean absolute deviation (MAD) and the root mean square error (RMSE) of the estimates of variance and covariance parameters for the SQD and the MSD relative to the complete data (the optimal design procedure improves efficiency and thus affects only variance and covariance estimates). The results are shown in Table 3.3. We obtain better values for the BIC- and KL- statistics and less missing information for the SQD as compared to the MSD. Parameter estimates are also closer to the true values for the SQD: the MAD is equal to 3.143 for 10 splits and 2.817 for 5 splits while these values are equal to 3.730 and 3.210 for the matrix sampling design. The missing information for the unconstrained split designs is 24% (ten splits) and 27% (five splits), and 22% and 29%, for constrained split designs, respectively, when we eliminate 50-60% of the questions. In contrast, the fraction of missing information for the MSD is consistently higher. Since

*Split Questionnaire Design*

these results support the performance of the SQD, we investigate its performance in an empirical setting in the next section.

Table 3.1: Candidate split set for a five block between-block design

<b>N<sub>s</sub></b>	<b>Block 1 Q1-4</b>	<b>Block 2 Q5-8</b>	<b>Block 3 Q9-12</b>	<b>Block 4 Q13-16</b>	<b>Block 5 Q17-20</b>
1	0	0	0	0	0
2	1	0	0	0	0
3	0	1	0	0	0
4	0	0	1	0	0
5	0	0	0	1	0
6	0	0	0	0	1
7	1	1	0	0	0
8	1	0	1	0	0
9	0	1	1	0	0
10	1	1	1	0	0
11	1	0	0	1	0
12	0	1	0	1	0
13	1	1	0	1	0
14	0	0	1	1	0
15	1	0	1	1	0
16	0	1	1	1	0
17	1	1	1	1	0
18	1	0	0	0	1
19	0	1	0	0	1
20	1	1	0	0	1
21	0	0	1	0	1
22	1	0	1	0	1
23	0	1	1	0	1
24	1	1	1	0	1
25	0	0	0	1	1
26	1	0	0	1	1
27	0	1	0	1	1
28	1	1	0	1	1
29	0	0	1	1	1
30	1	0	1	1	1
31	0	1	1	1	1
32	1	1	1	1	1

Note: The size of the restricted split is 26 by excluding the splits with indices 1 to 6.

Table 3.2: Performance of the modified Federov algorithm

K	# of Possible designs ( $N_D$ )	Complete Enumeration (sec.)	Modified Federov Algorithm (sec.) <sup>1</sup>
5 splits	65780	260	20
10 splits	5311735	10456	50
15 splits	7726160	13343	78

<sup>1</sup> The modified Federov Algorithm results are based on 1000 random starts.

Table 3.3: Simulation results for between-block designs

Design Measure	Unconstrained 10 Splits		Constrained 10 Splits		Unconstrained 5 Splits		Constrained 5 Splits	
	SQD <sup>a</sup>	MSD	SQD	MSD	SQD	MSD	SQD	MSD
MAD	3.143	3.73	2.471	2.773	2.817	3.21	3.001	3.102
RMSE	3.454	4.283	2.753	3.252	3.288	3.764	3.514	3.701
$\gamma^b$	0.243	0.317	0.217	0.284	0.269	0.306	0.294	0.304
% Missing	0.600	0.600	0.500	0.500	0.533	0.533	0.500	0.500
BIC	5232	7193	8777	8989	4570	8170	8764	8796
logL(D)	-2608	-3589	-4380	-4486	-2277	-4077	-4374	-4390

<sup>a</sup> SQD = Optimal Split Questionnaire Design, MSD= Matrix Sampling Design

<sup>b</sup>  $\gamma$  is the fraction of missing information

### 3.10 Empirical Data Application

We apply our procedure to a previously published empirical dataset obtained with the “Project 2000 Ninth Gvu Survey Web Attitude and Perceptions Questionnaire”<sup>9</sup>, which assesses how people use the Web and

<sup>9</sup> <http://elab.vanderbilt.edu/research/topics/flow/project2000.gvu9.htm>

### *Split Questionnaire Design*

their attitudes towards using it (Novak, Hoffman, and Yung, 2000). This type of survey, applied repeatedly to the same panel for the purpose of tracking consumer attitudes and behavior, may benefit from application of split questionnaire designs since it is conducted on a regular basis with an almost identical structure. Although this particular application is less than ideal to illustrate the performance of SQD, since the questionnaire is relatively short, we consider the use of a published questionnaire and publicly available data attractive. There are sixty-five questions, grouped into nine blocks according to content. The first block contains five questions about the role of the Web in life, the second block consists of eight questions on feelings while using the Web, the third block is composed of five questions related to Web activities, there are seven questions in the fourth block about perceptions on using the Web, the fifth block consists of seven questions about attitudes about using the Web, the sixth block contains eight questions about people feelings towards using the Web, the seventh and eighth block are comprised of ten and nine questions, respectively, about attitudes and perceptions, and the last block contains questions on flow and usage of Web information. The questions are assessed on 9-point Likert scales and are considered to be continuous and normally distributed for the purposes of the present study.

Data are available for two waves of the study conducted in two consecutive years. We use these as initialization and validation data, containing 500 and 1150 respondents, respectively. All data are complete. The advantage of having access to complete data is that it allows us to assess the performance of the SQD. A disadvantage of using such complete data is that we may underestimate the effect of the split

questionnaire design, since we do not benefit from the advantages of improved quality of the responses due to reduced respondent burden. Therefore, we also construct a field study with this questionnaire on which we report in the final part of this chapter. The initialization data are derived from the first wave of the survey, which we use for creating the split questionnaire. From the initialization data, we calculate the complete data parameter estimates. This enables us to obtain the design using the Federov algorithm to minimize the Kullback-Leibler distance. We investigate the following designs, where all designs in this study are constructed to be fully identified:

- a) Optimal split questionnaire (SQD) and matrix sampling designs (MSD),
- b) Designs with five or ten splits,
- c) Between-block and within-block designs,
- d) Unconstrained or constrained designs.

We consider the MSD (matrix sampling design) as a benchmark for the between-block design. For the within-block SQD, we use as benchmarks a random questionnaire design (RQD, in which questions within blocks are randomly assigned) as well as an ad-hoc procedure based on a principal components analysis of the items, as explained in more detail below. We use about the same total number of questions in all designs. We generate the MSD by randomly choosing five or ten splits from the candidate split matrix and evenly distributing them among respondents, eliminating responses from the complete data matrix  $Y$  according to the design in

### *Split Questionnaire Design*

question. For the RQD we apply the same procedure for each block separately, each time randomly selecting splits from the candidate split set. Since we have access to the complete data, we apply the constructed designs to those data to generate the missing data pattern. To compare the designs, we compute the KL distance and BIC statistics, the fraction of missing information, and MAD and RMSE, after imputing the missing data with the Gibbs sampler. We use informative priors obtained from the initialization data, for all designs. We run the Gibbs for 3000 iterations and save the last 600 draws from the predictive distribution for  $Y_{\text{mis}}$  as imputations; iteration plots show that the chains converge well before the end of the burn-in period.

#### **3.10.1 Between-Block Designs**

The MAD and RMSE measures shown in Table 3.4 reveal that the estimated parameters for the optimal SQD design are close to the complete data parameters. For both the five- and ten-split cases, the SQD improves significantly over the MSD, the MAD being 35% and 45% smaller respectively, and RMSE 34% and 45%. The improvement of the optimal designs over the currently used matrix sampling designs is substantial. The reason for the better performance of the five-split design, which results in 32% lower MAD and 31% lower RMSE than the ten-split design, is that the lower number of splits is associated with a smaller percentage of missing questions. For this particular application, the five-split optimal SQD results in a reduction of around 66% of the questions, with only a 14% information loss. With ten splits we obtain a greater reduction in the number of questions as compared to five splits. Here, while the SQD results in a 14%

loss of information, for the MSD the fraction of missing information is larger, 18%. The split questionnaires with five and ten splits are provided in Figure 3.3.

Table 3.4: Comparison of designs on empirical data

<b>BETWEEN-BLOCK DESIGNS</b>								
	Unconst. 10 Splits		Const. 10 Splits-5Blocks/Split		Unconst. 5 Splits		Const. 5 Splits-5 Blocks/Split	
	SQD	MSD	SQD	MSD	SQD	MSD	SQD	MSD
MAD	0.265	0.483	0.169	0.197	0.18	0.277	0.148	0.159
RMSE	0.378	0.682	0.24	0.319	0.262	0.399	0.203	0.215
$\gamma$	0.143	0.182	0.074	0.134	0.140	0.170	0.089	0.109
%Missing	0.735	0.735	0.492	0.492	0.662	0.662	0.440	0.440
BIC	18410	30298	57284	57655	15070	38740	64489	64675
logL(D)	-9195	-15139	-28631	-28817	-7525	-19360	-32234	-32327
<b>WITHIN-BLOCK DESIGNS</b>								
	10 splits			5 splits				
	SQD	RQD	PCA	SQD	RQD	PCA		
MAD	0.156	0.163	0.164	0.125	0.125	0.129		
RMSE	0.227	0.243	0.251	0.201	0.211	0.216		
$\gamma$	0.078	0.087	0.084	0.056	0.060	0.058		
%Missing	0.515	0.515	0.515	0.406	0.406	0.406		
BIC	44134	45186	45085	54890	55126	54979		
logL(D)	-22056	-22582	-22532	-27434	-27552	-27479		

<sup>a</sup> SQD = optimal Split Questionnaire Design, MSD= Matrix Sampling Design, RQD = Random Questionnaire Design, PCA = Principal Components Design

In addition, we investigate the case where constraints are imposed on the SQD. In particular, we construct designs in which each split consists of exactly five blocks. We choose this number, since we need at least five splits to generate fully identified designs under the constraint of five blocks per split. We repeat the design construction and imputation procedure on the empirical data, using five and ten splits, fixing each split to contain five

### *Split Questionnaire Design*

blocks. The results are given in Table 3.4. We focus first on the five-split design. In this case we reduce the number of questions with about 44%, while it was 66% for the unconstrained SQD. As a result, the constrained SQD yields 9% of missing information, while the unconstrained SQD yields 14% of missing information (these numbers are 7% and 14%, respectively, for the ten-split SQD). The fraction of missing information is also less for the constrained SQD than for the constrained MSD, as expected, but the  $\log L(D)$  and BIC for the constrained designs are worse than for the unconstrained designs. The RMSE and MAD measures reveal that the SQD estimates are close to those of the complete data, these measures are even smaller than for the unconstrained design. They are better than for the comparable MSD's, although the differences are smaller than for the unconstrained designs. The reason is that the constraints strongly limit the degrees of freedom for improvement over the MSD, since they reduce the size of the candidate split set. The optimal constrained five and ten-split designs are shown in Figure 3.4.

#### **3.10.2 Within-block Designs**

Using the prior estimates from the initialization data, we also construct optimal within-block designs by selecting questions within blocks, as described above. We compare the optimal SQD with designs in which the questions within blocks are selected randomly (RQD). To also compare to a stronger benchmark, we construct split designs using principal component analysis (PCA)<sup>10</sup>. We extract five and ten Varimax rotated components to

---

<sup>10</sup> We acknowledge an anonymous reviewer for this suggestion.

construct the splits. Questions in a block are discarded for a split if they contribute the least variance for that component. Every question was included at least once, and the design has the same number of questions as the SQD and RQD designs.

The results are shown in Table 3.4. We reduce 41% and 52% of the questions with the five- and ten- split within-block designs. The BIC and KL-distance of the optimal within-block designs are lower than the random design and the principal components design. The optimal within-block designs are also somewhat better in terms of RMSE and MAD of the parameter values, but the differences are not as large as for the between-block designs. The PCA designs are in between the RQD and optimal SQD on these measures. The average percentage of missing information is around 7.8% and 5.6% respectively for the optimal five- and ten-split designs. These numbers are better than for the corresponding random designs, with 8.7% and 6.0% respectively, and for the PCA designs, with 8.4% and 5.8%, respectively. The fraction of missing information for within-block designs, however, is substantially lower than for the between-block designs. MAD and RMSE of the five-split within-block designs are 31% and 23% lower than those of the between-block designs. For the ten-split designs they are 41% and 40% lower than those of the between-block design. However, the MAD and RMSE of the within-block designs are comparable to those of the constrained between-block designs. The optimal within-block designs are shown in Figure 3.5.

### *Split Questionnaire Design*

The estimates of the variances of the responses to the questions for the prior data, full and split questionnaires (after imputation) are shown in Table 3.5. As can be seen from the table, the prior estimates are close to complete questionnaire estimates of the current study. This illustrates the value of such prior estimates for the construction of split designs, but we further investigate the sensitivity of the optimal between- and within-block designs to these prior parameter values. For this purpose, randomly draw 50 sets of values from the sampling distribution of the parameters obtained from the initialization data and obtain optimal ten-split unconstrained and constrained between-block designs and within-block designs based on each of these sets. On average, we found 9.7 splits to be the same across these replications for the unconstrained between-block design<sup>11</sup>. For the constrained ten-split between-block design we find a lower average number of corresponding splits, 5.5. For the within-block design, on average only 2.2 splits were the same. Clearly, the within-block design is much more sensitive to the choice of the prior than the between-block designs. The size of the full candidate split set, as well as the use of the greedy design generating algorithm contribute to the high prior sensitivity of the within-block design. In particular, we find the sensitivity of the between-block design to the prior specification highly satisfactory.

---

<sup>11</sup> The maximum is 10, if all prior values yield exactly the same design, since there are ten splits in the design.

*Essays On Customization Applications in Marketing*

Table 3.5: Variance estimates after imputation<sup>1</sup>

	Full	Full	Between		Within		Full	Full	Between		Within
	Wave	Wave	Con.	Unc.	SQD		Wave	Wave	Con.	Unc.	SQD
	1	2	SQD	SQD	SQD		1	2	SQD	SQD	SQD
1	2.29	2.27	2.27	2.16	2.36	34	3.19	3.09	3.45	3.47	3.64
2	2.56	2.38	2.38	2.34	2.54	35	3.41	3.51	3.96	3.95	4.29
3	1.92	2.22	2.22	2.16	2.20	36	2.17	1.78	1.88	1.88	1.84
4	1.96	2.13	2.13	2.14	2.08	37	1.96	1.89	2.14	2.07	1.98
5	2.33	2.15	2.15	2.19	2.41	38	2.49	2.47	2.76	3.03	2.58
6	4.35	4.14	4.66	5.02	4.33	39	1.87	1.84	2.06	1.92	1.88
7	2.63	2.36	2.50	2.74	2.34	40	2.04	2.29	2.65	2.07	2.49
8	2.82	2.81	3.02	2.97	3.17	41	2.80	2.79	2.93	4.54	3.16
9	2.29	2.52	2.79	2.49	2.76	42	3.85	4.00	4.45	5.19	3.95
10	1.69	1.89	1.98	1.98	2.19	43	2.90	2.89	3.12	3.97	3.08
11	3.08	3.11	3.30	4.01	3.13	44	4.59	4.34	4.85	7.28	4.41
12	2.42	2.51	2.74	2.72	2.88	45	4.13	4.04	4.66	4.97	3.96
13	2.69	2.28	2.71	2.71	2.32	46	2.95	2.92	3.37	4.76	3.02
14	1.86	2.18	2.18	2.31	2.24	47	3.04	3.20	3.75	4.52	3.52
15	3.77	3.62	3.72	3.87	4.03	48	4.87	4.60	5.64	6.23	4.66
16	4.31	3.87	4.05	3.91	4.08	49	4.86	4.66	4.77	6.09	4.70
17	5.48	4.62	4.74	4.66	5.21	50	3.77	3.96	4.66	5.84	4.51
18	3.25	3.54	3.60	3.67	3.59	51	3.05	3.05	3.11	3.49	3.19
19	4.97	4.62	5.25	5.23	5.03	52	2.13	2.22	2.52	2.96	2.25
20	4.79	4.86	5.29	5.63	6.46	53	5.38	5.48	5.85	7.17	5.50
21	3.08	2.91	3.08	3.55	3.06	54	4.89	4.59	5.19	6.51	5.00
22	2.87	2.90	3.07	2.99	3.24	55	3.19	3.47	4.25	5.62	3.44
23	3.06	3.43	3.59	4.09	3.61	56	5.03	4.67	5.19	7.20	4.79
24	5.22	5.36	6.07	6.03	5.75	57	2.94	3.12	3.44	4.03	3.29
25	2.27	2.04	2.28	2.53	2.28	58	2.93	2.77	2.99	3.78	2.94
26	4.40	4.08	4.30	4.53	4.40	59	3.52	3.60	3.81	5.00	3.58
27	5.33	5.49	5.97	6.34	5.49	60	6.78	6.74	7.07	7.81	6.91
28	3.66	4.20	4.59	5.24	4.35	61	4.54	4.68	4.87	5.26	4.75
29	2.76	2.96	3.08	3.20	3.02	62	5.21	5.23	5.37	5.97	5.56
30	4.83	4.87	5.42	6.04	5.10	63	1.07	1.12	1.19	1.23	1.27
31	3.45	3.88	4.59	5.37	3.97	64	1.66	1.84	1.85	1.95	1.93
32	4.12	4.25	4.65	5.51	4.64	65	0.56	0.53	0.55	0.60	0.52
33	1.43	1.41	1.71	1.61	1.60						

<sup>1</sup> From the full questionnaire from the first and second wave survey, the constrained and unconstrained between- and within ten-split optimal designs

### **3.11 Field Study**

The above analysis illustrates that optimally designed split questionnaires can be beneficial, but only address that issue from a statistical perspective. In this section, we look into the behavioral issues of providing subjects split questionnaires. We conducted a field experiment to investigate whether with split questionnaires one may reduce boredom, fatigue, and completion time, which ultimately should increase the quality of data. We will also investigate respondents' attitudes towards the questionnaires, and assess whether using split questionnaires improves the reliability of constructs, compared to the full questionnaire.

For the field study, we use the exact questionnaire that was used in the empirical study above. We asked additional questions about boredom, which is scaled 1 (not at all bored) to 9 (extremely bored), fatigue which also is scaled 1 (not at all tired) to 9 (extremely tired). In addition, we assessed attitudes towards the questionnaire (three questions, five-point scale: repetitive-varied, very long-very short, boring-stimulating). We tested the full questionnaire, a ten-split between-block design, and a ten-split within-block design (see above) each on 63 subjects recruited from the subject pool from [withheld for confidentiality]. In total, 189 subjects responded to 21 versions of the questionnaire that were displayed on computer screens in the experimental lab. Computer aided questionnaires allowed us to record the exact time it took respondents to complete them. These average times to complete the full and split questionnaires differed significantly: 8 minutes for the complete questionnaire, and about 6 minutes

for each of the split questionnaires. This is a significant reduction of about 25% in completion time, with a 50% reduction in the number of questions. Note that even the full questionnaire with 65 questions can be completed relatively quickly --the longest it took any respondent was 10 minutes--, which makes it more difficult to identify the behavioral effects of the split questionnaires.

The mean scores for the scales are shown in Table 3.6. A MANOVA across all measures reveals a significant difference between the complete and between-block design ( $p < 0.01$ ) and the complete and within-block design ( $p < 0.01$ ), but not between the latter two. The mean boredom score for the full questionnaire is 5.44, which is significantly higher than that for the within-block questionnaire, which is 4.98. The differences with the between-block design, which has an intermediate boredom score of 5.23, are not significant. This may indicate that feelings of boredom are primarily caused by repetition of the relatively similar questions within blocks, which occurs less in the within-block design. The respondents that completed the full questionnaire report feeling more tired than those receiving the between-block design, the mean scores being 4.32 and 3.57. The within-block tiredness score is intermediate, 3.73, and not significantly different from the other two. This may point to feelings of tiredness being more strongly associated with switching between different topics, which occurs less often in the between-block design due to a reduction of the number blocks. The split questionnaire designs are evaluated more favorably than the complete questionnaire, the between- and within-block designs being

### *Split Questionnaire Design*

seen as less repetitive (5.32 and 4.20 versus 5.68) and less boring (4.77 and 4.42 versus 4.94) than the complete questionnaire. The scores for the within-block design are significantly better than those for the between-block design. The within-block design is also considered to be significantly less long than the complete questionnaire design (3.13 versus 3.68; and 3.54 for the between-block design, which is not significantly different from the former two). The shorter perceived duration of the within-block design may be associated with its lower perceived boredom discussed above, since its actual duration is about 20 seconds longer than that of the between-block design (the longer duration may have to do with the reading and processing of the separate instructions for each block).

Table 3.6: Item means and SDs from the field experiment

	<b>FULL QUESTIONNAIRE</b>	<b>BETWEEN- BLOCK SQD</b>	<b>WITHIN- BLOCK SQD</b>
Duration	476.92 (95.01)	344.48 <sup>a1</sup> (146.552)	364.02 <sup>b1</sup> (93.57)
Boredom	5.44 (2.09)	5.23 (1.95)	4.98 <sup>b1</sup> (2.00)
Fatigue	4.32 (2.55)	3.57 <sup>a2</sup> (2.27)	3.73 (2.02)
Repetitive	5.68 (1.37)	5.32 <sup>c1</sup> (1.22)	4.70 <sup>b1c1</sup> (1.78)
Long	3.68 (1.54)	3.54 (1.56)	3.13 <sup>b1</sup> (1.25)
Boring	4.94 (1.28)	4.77 <sup>c1</sup> (1.11)	4.42 <sup>b1c1</sup> (1.25)
Cronbach's alpha	0.66	0.66	0.67
Item Variance	3.34	2.36 <sup>a1</sup>	2.30 <sup>b1</sup>

Notes: The values in parenthesis are standard deviations. N=189. Duration mean values are in seconds. Superscripts indicate the significance of the differences between means of the full & between- (<sup>a</sup>), full & within- (<sup>b</sup>) and between- & within- (<sup>c</sup>) block questionnaires; <sup>1</sup> p=0.05 , <sup>2</sup> p=0.10

In short, split questionnaire designs decrease completion time, fatigue, boredom and non-response and are evaluated more positively by respondents, where it seems that the within-block design has a somewhat more favorable behavioral effect than the between-block design. These effects may impact the quality of the data. For each of the three questionnaires, respondents could skip every question displayed on the screen. There were 33 skip-responses for the full questionnaire, 7 for the

### *Split Questionnaire Design*

between- and 5 for the within-block design. These responses start only after the first twelve questions and mostly occur in the last half of the questionnaires. This indicates that the use of split questionnaires may substantially reduce item non-response. Second, the effect of the questionnaire design on the average item variances and Cronbach's alpha were investigated. The questionnaire consists of 13 constructs that are each measured with several items. There were no statistically significant differences in the average Cronbach's alpha, estimated after multiple imputation of the missing data of the between- and within-block split questionnaire designs. However, we did find significant differences in item variances between the full- and split questionnaire designs. The differences between between-block and within-block designs are not significant. The average item variance for the full questionnaire is 3.34, which is significantly higher than for the between-block design, with 2.36, and the within-block design, 2.30. This means that subjects who answered the questions in the within-block or between-block designs responded to the items that measure the same construct more consistently. Thus, the quality of the data we obtained from the between-and within-block split questionnaire designs tends to be better than that of the full questionnaire. Again, we note that with a maximum average completion time of eight minutes, the complete questionnaire is relatively short. For longer questionnaires, the effects may be even larger.

### **3.12 Conclusion**

Split questionnaires present opportunities for application in consumer panels, offering the potential to obtain higher quality information from respondents faster and at a substantially lower cost. In this chapter, we first propose a methodology to split questionnaires optimally into sub-components with minimal information loss by applying optimal experimental design methods. We proposed the Kullback-Leibler distance as a design criterion, applied the modified Federov algorithm to search over the design space, and illustrated that good designs can be constructed rapidly in spite of the demanding task. Split questionnaire designs were shown to have desirable statistical and behavioral properties, relative to complete questionnaires or questionnaires constructed with ad-hoc methods.

We have investigated two different types of split questionnaire designs based on the contextual placement of questions in blocks. The first method, producing between-block designs, places blocks as a whole into different split versions of the questionnaire. Optimizing the allocation of the blocks across the splits is a much more feasible task than allocating individual questions to splits. Additional constraints, such as on the number of blocks per split, can easily be accommodated and may further reduce the number of questions asked from each respondent. Between-block designs result in estimates close to those obtained from the complete data, reducing completion time and respondent fatigue. The second method, producing within-block designs, is based on choosing questions in each block. For

### *Split Questionnaire Design*

these designs, the optimization task is very demanding, so that we needed to use a greedy algorithm to find the optimal design. As a consequence, the within-block designs are not strictly optimal, nor can they easily be constructed to be fully identified. However, they do provide improved efficiency, yielding parameter estimates that are closer to the complete data estimates and less missing information than designs constructed with heuristic procedures. Their performance in terms of parameter estimates and missing information tends to be better than that of the between-block designs, but they are substantially more sensitive to the values of the prior estimates.

Our field study shows that the behavioral reaction of respondents to split questionnaires is more favorable than to the complete questionnaire, in terms of duration, boredom, and fatigue, amongst others. The response to within-block designs tends to be more positive than that to a between-block design, since respondents feel less bored, and perceive the questionnaire as less long, boring and repetitive. The between-block design, however, results in less respondent fatigue. The choice between the within-block and between-block designs may therefore be based on either statistical or behavioral criteria. From our investigation, it appears that the between-block design has better statistical properties, since it is feasible to construct fully identified designs with little sensitivity to the prior estimates. However, the within-block design still performs quite satisfactorily, yields parameter estimates comparable to constrained between-block designs, and elicits a more positive reaction from respondents. However, the high sensitivity of these designs to prior estimates warrants further study.

The validity of the prior knowledge when constructing the split questionnaire design is an important issue. Whereas prior knowledge can be easily obtained in panel or tracking surveys conducted on a regular basis with almost identical questions and blocks, it may be less easy to obtain in other settings. In those cases, subjective prior distributions for the model parameters can be assessed, which in many cases would involve the elicitation of priors from consumers, decision makers or other subject-matter experts. Chaloner (1996) provides an overview of the various approaches to elicitation based on the ways people think about and update probabilistic statements. It is of interest to consider prior uncertainty on the parameters in constructing the designs, and to construct designs integrating the design criterion over the prior distribution of the parameters (Sandor and Wedel, 2001). This may in particular be worthwhile for within-block designs, which were revealed to have high sensitivity to the prior specification. For between-block designs, in particular in panel data applications such as the one presented above, this may not be needed, since the prior parameter values can be fairly precisely estimated from the available pilot data, and the designs themselves were shown to be quite insensitive to the prior parameter values. We leave these issues for future research.

### 3.13 Appendix

#### 3.13.1 KL-Distance for Mixed Data

The incomplete data log-likelihood of mixed data is derived below using the general location model (Olkin and Tate, 1961; Krzanowski, 1983). We have the data matrix  $Y_{N \times (p+q)} = (X, Z)$ , where  $X = (X_1, \dots, X_p)'$  and  $Z = (Z_1, \dots, Z_q)$  represent the continuous and categorical variables, respectively. Each column variable in  $Z$ ,  $z_j$  has  $c_j$  levels, and these categorical variables form a  $q$ -dimensional contingency table with a total number of cells  $C = \prod_j^q c_j$ . The frequencies in this table are contained in  $W = (w_{c_1}, w_{c_2}, \dots, w_{c_q})$ . The marginal distribution of the categorical variable  $Z$  is multinomial distribution  $(w | \pi = (\pi_1, \pi_2, \dots, \pi_c)') \sim \text{multinomial}(\pi)$  with  $\sum_{i=1}^c \pi_i = 1$  and the conditional distribution of the continuous variables  $X$  given categorical variables  $Z$  (i.e. given a particular cell) is multivariate normal with different means across the cells defined by the categorical variables, but with a common covariance matrix  $(X | Z = w, \mu_i, \Sigma \sim N(\mu_i, \Sigma))$ , where  $\mu_i$  is the mean of  $X$  in the cell specified by  $z$ , and  $\Sigma$  is the common conditional covariance of  $X$  across cells of the contingency table). The KL-distance in this case reduces to the

incomplete-data log-likelihood:

$$\begin{aligned}
 L(\mu_z, \Sigma, \pi | D) &= \sum_{i=1}^N \log f(x_{i,obs} | C_i, \mu_z, \Sigma) + \sum_{c \in C_i} \log f(\pi_c) \\
 &= -\frac{1}{2} \sum_{i=1}^N p_i \log(2\pi) - \frac{1}{2} \sum_{i=1}^N \log |\Sigma_{i,obs}| \\
 &\quad + \sum_{i=1}^N \log \left[ \sum_{c \in C_i} \pi_c \exp \left\{ -\frac{1}{2} (X_{i,obs} - \mu_{i,obs,c})' \Sigma_{i,obs}^{-1} (X_{i,obs} - \mu_{i,obs,c}) \right\} \right] \quad (A1)
 \end{aligned}$$

where  $X_{i,obs} = X_{ij}d_{ij}$ , where  $d_{ij}$  is the element of design matrix D.

### 3.13.2 Missing Information Principle

Assume that  $f(Y|\theta)$  is the probability distribution of  $Y = (Y_{\text{obs}}, Y_{\text{mis}})$  and parameter  $\theta$ . The distribution of the complete data  $Y$  can be factored with  $f(Y_{\text{obs}}|\theta)$ , the density of the observed data  $Y_{\text{obs}}$ , and  $f(Y_{\text{mis}}|Y_{\text{obs}},\theta)$ , the density of the missing data given the observed data, is represented as

$$f(Y|\theta) = f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) = f(Y_{\text{obs}}|\theta)f(Y_{\text{mis}}|Y_{\text{obs}},\theta) \quad (\text{B.1})$$

The decomposition of the loglikelihood that corresponds to (B.1) is

$$\ell(\theta|Y) = \ell(\theta|Y_{\text{obs}}, Y_{\text{mis}}) = \ell(\theta|Y_{\text{obs}}) + \ell(Y_{\text{mis}}|Y_{\text{obs}},\theta) \quad (\text{B.2})$$

Since directly maximizing the incomplete-data likelihood  $l(\theta|Y_{\text{obs}})$  with respect to  $\theta$  for fixed  $Y_{\text{obs}}$  to estimate  $\theta$  can be difficult, we can write B.3 with the observed loglikelihood  $l(\theta|Y_{\text{obs}})$ , the complete-data loglikelihood  $l(\theta|Y)$ , and the missing part of the complete-data loglikelihood  $l(Y_{\text{mis}}|Y_{\text{obs}},\theta)$

$$\ell(\theta|Y_{\text{obs}}) = \ell(\theta|Y) - \ell(Y_{\text{mis}}|Y_{\text{obs}},\theta) \quad (\text{B.3})$$

The observed information matrix  $l(\theta|Y_{\text{obs}})$  can be found directly by differentiating the loglikelihood  $l(\theta|Y_{\text{obs}})$  twice with respect to  $\theta$ . Alternatively, differentiating  $l(\theta|Y_{\text{obs}})$  twice with respect to  $\theta$  yields for any  $Y_{\text{mis}}$

$$\ell(\theta|Y_{\text{obs}}) = \ell(\theta|Y_{\text{obs}}, Y_{\text{mis}}) + \frac{\partial^2 \ln f(Y_{\text{mis}}|Y_{\text{obs}},\theta)}{\partial \theta \partial \theta}, \quad (\text{B.4})$$

where  $I(\theta|Y_{\text{obs}}, Y_{\text{mis}})$  is the observed information based on  $Y=(Y_{\text{obs}}, Y_{\text{mis}})$  and the negative of the last term is the missing information from  $Y_{\text{mis}}$ . Taking expectations over the distribution of  $Y_{\text{mis}}$  given  $Y_{\text{obs}}$  and  $\theta$  yields

$$\text{Observed Information} = \text{Complete Information} - \text{Missing Information} \quad (\text{B.5})$$

The observed information equals the (conditional expected) complete information minus the missing information. This has been called the missing information principle by Orchard and Woodbury (1972).

The decomposition of the observed information is particularly simple in case the complete data  $Y$  have a distribution from the regular exponential family defined by

$$f(Y|\theta) = b(Y)\exp(s(Y)\theta)/a(\theta) \quad (\text{B.6})$$

where  $\theta$  denotes a  $(d \times 1)$  parameter vector,  $s(Y)$  denotes a  $(1 \times d)$  vector of complete-data sufficient statistics, and  $a$  and  $b$  are functions of  $\theta$  and  $Y$ , respectively. The complete information is  $\text{Var}(s(Y)|\theta)$ , and the missing information is  $\text{Var}(s(Y)|Y_{\text{obs}}, \theta)$ . Thus the observed information is the difference between the unconditional and conditional variance of the complete-data sufficient statistic.

$$I(\theta|Y_{\text{obs}}) = \text{Var}(s(Y)|\theta) - \text{Var}(s(Y)|Y_{\text{obs}}, \theta), \quad (\text{B.7})$$

In sum, according to the missing data information principle, the missing information is equal to the variance difference between the complete data and the incomplete data. In questionnaire design, the missing information measures the increase in variance of estimation due to nonresponse, and

### *Split Questionnaire Design*

is determined by response rates and the ability of observed values to predict missing values successfully.

### **3.13.3 Figures**

#### **Description of Blocks:**

Block 1: Five questions about the role of the Web in life.

Block 2: Eight questions about the feeling while using the Web

Block 3: Five questions related to the Web activities feeling while using the Web

Block 4: Seven questions about and perceptions on using the Web

Block 5: Seven questions about attitudes and perceptions on using the Web

Block 6: Eight questions about peoples' feelings towards using the Web

Block 7: Ten questions on attitudes and perceptions

Block 8: Nine questions about attitudes and perceptions on using the Web

Block 9: Six questions about flow and usage of the web.

*Split Questionnaire Design*

Figure 3.3: Optimal Unconstrained Between-Block Designs for the Empirical Data

THE OPTIMAL 10-SPLIT UNCONSTRAINED BETWEEN-BLOCK SQD

Resp.No.	Block 1 Q1-5	Block 2 Q6-13	Block 3 Q14-18	Block 4 Q19-25	Block 5 Q26-31	Block 6 Q32-40	Block 7 Q41-50	Block 8 Q51-59	Block 9 Q60-65
1-115									
116-230									
231-345									
346-460									
461-575									
576-690									
691-805									
806-920									
921-1035									
1036-1150									

THE OPTIMAL 5-SPLIT UNCONSTRAINED BETWEEN-BLOCK SQD

Resp.No.	Block 1 Q1-5	Block 2 Q6-13	Block 3 Q14-18	Block 4 Q19-25	Block 5 Q26-31	Block 6 Q32-40	Block 7 Q41-50	Block 8 Q51-59	Block 9 Q60-65
1-230									
231-460									
461-690									
691-920									
921-1150									

Note: shaded are observed, blank are missing blocks.

Figure 3.4: Optimal Constrained Between-Block Designs for the Empirical Data

THE OPTIMAL 10-SPLIT 5-BLOCK BETWEEN-BLOCK SQD

Resp.No.	Block 1 Q1-5	Block 2 Q6-13	Block 3 Q14-18	Block 4 Q19-25	Block 5 Q26-31	Block 6 Q32-40	Block 7 Q41-50	Block 8 Q51-59	Block 9 Q60-65
1-115									
116-230									
231-345									
346-460									
461-575									
576-690									
691-805									
806-920									
921-1035									
1036-1150									

THE OPTIMAL 10-SPLIT 5-BLOCK BETWEEN-BLOCK SQD

Resp.No.	Block 1 Q1-5	Block 2 Q6-13	Block 3 Q14-18	Block 4 Q19-25	Block 5 Q26-31	Block 6 Q32-40	Block 7 Q41-50	Block 8 Q51-59	Block 9 Q60-65
1-230									
231-460									
461-690									
691-920									
921-1150									

Note: shaded are observed, blank are missing blocks.

*Split Questionnaire Design*

Figure 3.5: Optimal Within-Block Designs for the Empirical Data

THE OPTIMAL 10-SPLIT WITHIN-BLOCK SQD

<b>Bl. 1</b> <b>Q1-5</b>	<b>Bl. 2</b> <b>Q6-13</b>	<b>Bl. 3</b> <b>Q14-18</b>	<b>Bl. 4</b> <b>Q19-25</b>	<b>Bl. 5</b> <b>Q26-31</b>	<b>Bl. 6</b> <b>Q32-40</b>	<b>Bl. 7</b> <b>Q41-50</b>	<b>Bl. 8</b> <b>Q51-59</b>	<b>Bl. 9</b> <b>Q60-65</b>
00110	00000101	00101	1100000	0011101	00011111	0110001100	01111110	111010
11111	11111111	11111	1000100	1101010	01000010	0011100100	11100011	011111
00011	10000001	10100	1010000	0101010	00010001	1100110100	01101111	100010
10010	01000100	00101	0010001	0111110	11011110	0111110010	10011100	110010
10100	01010000	00101	1000010	1001001	10101100	1110111010	11001101	010011
01100	10010000	00011	0010010	1101110	10010010	0100011110	01011101	001101
00101	00101000	11000	1000010	1110101	00011100	1001111101	00101100	110101
11000	10010000	10010	0010001	1100110	01011110	1011011011	11001110	010001
01001	01000100	00011	0011000	0010011	11100011	1110010111	01100000	110111
10001	00110000	10100	0001001	0111100	10110111	0000000011	10101101	110011

THE OPTIMAL 5-SPLIT WITHIN-BLOCK SQD

<b>Block 1</b> <b>Q1-5</b>	<b>Block 2</b> <b>Q6-13</b>	<b>Block 3</b> <b>Q14-18</b>	<b>Block 4</b> <b>Q19-25</b>	<b>Block 5</b> <b>Q26-31</b>	<b>Block 6</b> <b>Q32-40</b>	<b>Block 7</b> <b>Q41-50</b>	<b>Block 8</b> <b>Q51-59</b>	<b>Block 9</b> <b>Q60-65</b>
00110	01101101	00110	1010110	1111111	01100100	1110011101	00010101	101011
10010	10000101	11111	0101100	1101110	11110111	0000110011	01111110	011101
11111	00100011	01110	1011110	1101100	11011000	0111101011	11111111	100111
10100	01001111	00011	1101001	1001101	11101111	1101101110	10000110	110010
00101	11111111	10001	0110111	0101010	11001010	0001011001	00011100	001011

*Essays On Customization Applications in Marketing*



## **Chapter 4**

# **Promotion Customization across Multiple Categories**

### **4.1 Introduction**

In the previous chapter, split questionnaire designs have been proposed as a means of efficiently collecting data and a methodology was developed to design optimal split questionnaires based on prior information, which is obtained using full questionnaires from pilot or past studies. In this chapter, we apply the Bayesian framework to customize promotions across multiple categories. The purpose is to identify the optimal subset of categories to promote to each individual, based on associations in purchase behavior across categories.

As a consequence of technological developments, especially with the explosion in the usage of the World Wide Web, it is now possible to acquire immediate information about consumers and to meet their needs by customizing products and product-related information in real time. Nowadays, more and more companies use such customized and targeted online promotion programs, including customization of ads, email-messages, sales-promotions to loyalty card users, electronic coupons etc. Recently, Bucklin et al. (2002) proposed customization and automation as

key areas of research in e-commerce. Targeting and customization issues have long been of interest in marketing. For example, Rossi et al. (1996) have applied targeting to a coupon delivery problem. Researchers such as Ansari and Mela (2003), and Bertsimas and Mersereau (2003) have examined the customization of information by means of e-mail marketing content on the internet; Gooley and Lattin (2000) have studied customization of advertisement content; Zhang and Krishnamurthi (2004) have shown how to customize online price promotions and Ansari, Essagaier and Kohli (2000) have analyzed customization of product offers.

Online retailing has increased its revenues by almost \$72 million. This demonstrates that online retailing has become a major force in the consumer industry. Repeat buyers account for more than half of online revenues, and stimulating repeat-buying renders online retailing more efficient and therefore more profitable. Because of the information available on repeat customers, online retailers can customize their offers to them. This improves conversion rates and may justify higher margins ([www.bcg.com/publications](http://www.bcg.com/publications)). Retailers therefore are continuing to invest in tools to facilitate repeat-buying, promotional tools being chief among these. The dynamic character of the Internet makes relatively easy for retailers to offer promotions to individual customers “on the fly” to guide their current decisions, by using information from their previous choices. Hence, customized delivery of promotions via email or on the web is becoming vital in online retailers strategies.

### *Promotion Customization across Multiple Categories*

Online grocery stores such as Peapod ([www.peapod.com](http://www.peapod.com)) and NetGrocer ([www.netgrocer.com](http://www.netgrocer.com)) use various customization services for the grocery shopping process on the Internet. Such services involve creating personal lists for products (that frequently purchased, for weekend parties, or for special occasions), creating lists of the items available in a customer's pantry and refrigerator, and then suggesting recipes in which those items can be used. Another example of an advanced customized promotion program is that employed by CVS Pharmacy ([www.cvs.com](http://www.cvs.com)). CVS uses loyalty cards to offer different sales-promotions for low-tier, middle-tier, and top-tier customers. It also uses target mailings with segment-level content and customized offers. Moreover, CVS customizes promotions at the cash register using previous category purchase histories. Price promotions and coupons are an important part of any customization process.

According to the Association of Coupon Professionals, Internet-delivered coupons, although still a controversial topic in the industry, saw a five-fold increase in distribution as entrepreneurial marketers sought better ways to target and deliver effective incentives ([www.couponpros.org](http://www.couponpros.org)). Merchants use information about their shoppers to target e-coupons from these three possible types of data. The first one is shoppers' socio-demographic profiles, which can be obtained from external sources or directly obtained from shoppers in the form of answers to questionnaires and surveys. The second is the shopper's clickstream, which is a raw log of the web pages requested by the shopper in the merchant's store. The last one is shoppers' transactions, such as items purchased, the recipient of the purchases, items added to a shopping cart, and e-coupons offered,

accepted and used. Customized coupon offers can be delivered through e-mail or on the web. The e-coupon is a short piece of text that can carry a short message. Promotional web pages show a certain number of coupon offers for different products and categories. Electronic distribution of coupons has become more widespread under programs such as Catalina Marketing Incorporated's (CMI; [www.catalinamarketing.com](http://www.catalinamarketing.com)) checkout coupon and frequent shopper programs, in which households receive in store coupons and/or volume discounts through the Internet ([www.valupage.com/Entry.pst](http://www.valupage.com/Entry.pst)). In addition, specialized web-based promotion companies offer promotions for a range of products and categories for different manufacturers and retailers ([www.dealcatcher.com](http://www.dealcatcher.com), [www.coolsavings.com](http://www.coolsavings.com), [www.allonlinecoupons.com](http://www.allonlinecoupons.com), [www.findsavings.com](http://www.findsavings.com)).

A limitation of the customized promotion programs in practice as well as those described in the academic literature to date is that electronic coupons are issued by companies based on customer information in only one category at a time. In this chapter, we consider the development of a customization method by focusing on the selection of what product categories to promote, across multiple product categories, taking into account the dependencies in consumers' purchase behavior across those categories. Our method can be applied to, for example, checkout coupons printed on cash register receipts or e-coupon promotions delivered on web pages or through e-mail. In practice, these receipts, web pages and e-mails can show only a limited number of coupon offers, mostly due to space limitations, and our purpose is to select the most suitable categories to

### *Promotion Customization across Multiple Categories*

promote to individuals to minimize the customer effort, as well as maximize the retailer revenue.

Our approach obtains cross-category purchase incidence and expenditure information and analyzes this as input to develop customized coupon programs. Our promotions design problem is to determine which category from many possible categories to assign to a customer for promotion. The approach has two stages. First of all, we obtain prior information about consumer choices, price and the promotional environment in online retail stores, at the category level. We obtain current and past purchase history information from online transaction data. We construct a consumer response model using total expenditure and multicategory purchase incidence information, while dealing with consumer heterogeneity. Our flexible model of heterogeneity accommodates observable and unobservable heterogeneity and produces household level inferences for targeting purposes. We consider the interdependence in consumer purchases among multiple categories (see e.g., Manchanda, Ansari, and Gupta, 1999). We model the incidence decisions and expenditures jointly. We use the expected expenditures on a shopping trip if we promote a category or not as a criterion to find the optimal promotion design, i.e. our design maximizes consumer expenditure, and thus the firm's revenue. The promotion design problem, however, increases exponentially with the number of categories. At the second stage, therefore, using estimates of this model, we search all possible category-promotions designs with the modified Federov algorithm and determine the optimal design, which gives maximum expenditures on a shopping basket to maximize revenue of online retailers.

Our objectives in this chapter are summarized as follows:

- Developing a joint heterogeneous model of category purchase incidence and expenditure decisions across multiple categories,
- Using this category buying behavior information, we develop a method for designing an optimal customized promotion plan, specifying what categories to promote to which consumers.

## **4.2 Literature Review**

We model consumer responses to promotions considering purchase incidence and expenditure together to obtain knowledge about correlations across categories. Several previous studies in the marketing literature have considered promotion effects at the multicategory level. These studies have been developed and estimated in three different ways: 1) for one product category and for a specific purchase variable at a time, 2) for three purchase variables (purchase incidence, brand choice and purchase quantity) simultaneously within a single product category, 3) for multiple purchase outcome variables simultaneously across multiple product categories. For example, Bolton (1989) finds that the effects of category display and feature activity are much larger than the effects of brand prices, display, and feature activity. While Gupta (1988) models customers' brand choice, inter-purchase time and purchase quantity decisions separately, he uses category level marketing-mix variables in the inter-purchase time and purchase quantity models. Seetharaman, Ainslie and Chintagunta (1999)

### *Promotion Customization across Multiple Categories*

study category level household state dependence considering brand choice behavior across five product categories. In this paper, they investigated whether households exhibit similar sensitivities to the marketing mix variables in different product categories. Fader and Lodish (1990) use IRI Marketing Factbook data from 331 product categories to explore the relationship between category structure (e.g. purchase cycle, penetration, etc.) and promotional movement (e.g. volume sold on price cuts, display and feature, etc.). They report systematic relationships between category characteristics and the effect of promotional policies. Narasimhan et al. (1996) study the relationship between product category characteristics and promotional elasticity using data from 108 product categories. They consider three types of promotions (regular price cuts, features, and displays) and seven category characteristics. They report that promotions obtain the highest response for brands in easily stockpiled, high penetration categories with short purchase cycles. Whereas the latter two studies ignore interdependence in consumer's purchases across multiple categories, Mulhern and Leone (1991), Chintagunta and Haldar (1998), and Manchanda and Gupta (1997) explicitly allow for dependency across multi-category items. Ainslie and Rossi (1998) measure the covariance of both observed (linked to measured characteristics of households) and unobserved heterogeneity in marketing sensitivity across two categories. Their focus is on the measurement of cross-category correlations in conditional choice behavior. Chib et al. (2002) modeled and estimated the purchase incidence model for twelve categories. In this paper, they illustrated disregarding cross-correlations across multiple categories in shopping basket models causes underestimation of the magnitude of cross-

category correlations and overestimation of the effectiveness of the marketing mix, and additionally ignoring unobserved individual heterogeneity results in overestimation of cross-category correlations and underestimation of the effectiveness of the marketing mix. Our model in this chapter is in the third class i.e. multiple purchase outcome variables are estimated simultaneously across multiple product categories. As far as we know, there is no previous study on joint modelling of purchase incidence and expenditure across multiple categories in the marketing literature and our study differs from the previous studies due to the consideration of correlations between purchase incidence and expenditure in the model estimation.

### **4.3 Methodology**

Our approach consists of two connected stages. At the first stage, we construct a consumer response model to obtain information about cross-category promotion effects, considering heterogeneity across households. At the second stage jointly executed with the first, we choose for every individual a limited set of promotions from all possible promotions with the modified Federov algorithm, using total shopping expenditure as a criterion.

#### **4.3.1 The Model**

Consumer purchase of multiple categories in a shopping trip can be characterized in terms of two related decisions: which categories to choose, and how much to spend. The role of the price and promotion variables in

the purchase process at the category level makes the joint modeling of the incidence decision and the expenditure decision on a shopping trip necessary. Modeling these two decisions separately, that is, using for example a multivariate probit model for category incidence and a regression model for the expenditures, is incorrect and yields biased estimates if expenditure decisions are not independent of the category incidence decisions. For this reason, we use a censored regression (tobit) model. A similar approach was previously used by Krishnamurthi and Raj (1988) to jointly model purchase quantity and brand-choices. We will use a hierarchical multivariate type-2 tobit to jointly model category choice incidence and expenditure (previous applications of the tobit models in marketing include those by DeSarbo and Choi (1999), for modeling consumer search behavior; by DeSarbo and Jedidi (1995) in a consideration set application; and by Bucklin and Sismeiro (2003) to model web site browsing behavior). The details of the model are explained below.

#### **4.3.2 Consumer Response Model**

We investigate the predictive relationship of covariates with category-incidences and expenditures through a censored regression model describing category incidence and expenditures simultaneously. Let us assume  $H$  households, represented by index  $h=1,\dots,H$ , making purchase incidence decisions across a set of  $J$  product categories,  $j=1,\dots,J$ , on a total of  $T$  shopping trips, indexed by  $t=1,\dots,T$ .  $Y_{2ht} = [Y_{2h1t}, Y_{2h2t}, \dots, Y_{2hjt}]$  are binary dependent variables with consumer's product-category incidence decision outcomes.  $Y_{1ht} = [Y_{1h1t}, Y_{1h2t}, \dots, Y_{1hjt}]$  is our expenditure vector. We

denote the predictor variables (marketing policy variables comprise price and promotion), as  $X_{ht} = [X_{h1t}, X_{h2t}, \dots, X_{hjt}]$ . The observed choice vector (incidence of category-choice),  $Y_2$  is

$$Y_{2hjt} = \begin{cases} 1, & \text{if } Y_{2hjt}^* \geq 0 \\ 0, & \text{if } Y_{2hjt}^* < 0 \end{cases} \quad (4.1)$$

Expenditures are observed only when the indicator variable for category choice,  $Y_{2hjt}$ , takes on the value 1:

$$Y_{1hjt} = \begin{cases} Y_{1hjt}^* & \text{if } Y_{2hjt} = 1 \\ 0 & \text{otherwise} \end{cases} \quad (4.2)$$

We can write the models for the latent utilities,  $Y_{2j}^*$ , and the logarithm of the latent expenditure variable,  $Y_{1j}^*$ , for the J categories as:

$$\begin{bmatrix} Y_{1h1}^* \\ Y_{1h2}^* \\ \dots \\ Y_{1hJ}^* \end{bmatrix} = \begin{bmatrix} X_{1h1} & 0 & \dots & 0 \\ \dots & X_{1h2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & X_{1hJ} \end{bmatrix} \begin{bmatrix} \beta_{1h1} \\ \beta_{1h2} \\ \dots \\ \beta_{1hJ} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1h1} \\ \varepsilon_{1h2} \\ \dots \\ \varepsilon_{1hJ} \end{bmatrix} \quad (4.3a)$$

$$\begin{bmatrix} Y_{2h1}^* \\ Y_{2h2}^* \\ \dots \\ Y_{2hJ}^* \end{bmatrix} = \begin{bmatrix} X_{2h1} & 0 & \dots & 0 \\ \dots & X_{2h2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & X_{2hJ} \end{bmatrix} \begin{bmatrix} \beta_{2h1} \\ \beta_{2h2} \\ \dots \\ \beta_{h2J} \end{bmatrix} + \begin{bmatrix} \varepsilon_{2h1} \\ \varepsilon_{h22} \\ \dots \\ \varepsilon_{2hJ} \end{bmatrix} \quad (4.3b)$$

*Promotion Customization across Multiple Categories*

Each of the  $Y_{h1}^*, Y_{h2}^*, \dots, Y_{hj}^*$  and  $\varepsilon_{h1}, \varepsilon_{h2}, \dots, \varepsilon_{hj}$  are  $T \times 1$  vectors. The matrices  $X_{hj}$  are of order  $T \times k$  and the vectors  $\beta_{hj}$  are of order  $k \times 1$ , with  $k$  the number of predictor variables.  $X$  is the matrix of category price, category promotion and intercept variables and has dimensions  $T \times kJ$ . Thus,  $\varepsilon_{1j}$ ,  $j=1, \dots, J$ , is an  $HT$  vector of error terms such that  $E(\varepsilon_{1j})=0$  and  $E(\varepsilon_{1i}\varepsilon'_{1j}) = \sigma_{1ij}I_{HT}$   $i, j=1, 2, \dots, J$  and  $\varepsilon_{2j}$ ,  $j=1, \dots, J$ , is an  $HT$  vector of error terms such that  $E(\varepsilon_{2i})=0$  and  $E(\varepsilon_{2i}\varepsilon'_{2j}) = \sigma_{2ij}I_{HT}$   $i, j=1, 2, \dots, J$ . Further, we have  $E(\varepsilon_{1i}\varepsilon'_{2j}) = \sigma_{12ij}I_{HT}$ .

We can re-write our equations as:

$$Y_{1ht}^* = X'_{ht}\beta_{1h} + \varepsilon_{1ht} \quad (4.4a)$$

$$Y_{2ht}^* = X'_{ht}\beta_{2h} + \varepsilon_{2ht} \quad (4.4b)$$

where the  $j$ -th row of the matrix  $X_{ht}$  contains all explanatory variables including price and promotion, and the intercepts for each product category, influencing the utility and expenditure of the  $j$ -th category. The error terms are  $\varepsilon_{ht} = [\varepsilon_{h1t}, \varepsilon_{h2t}, \dots, \varepsilon_{hjt}]$ . As unobserved factors informing the utilities may be common across categories, we assume that  $\varepsilon_{2ht} \sim MVN[0, \Sigma_{2(J \times J)}]$ , where  $\Sigma$  is a  $J \times J$  covariance matrix. Coincidence captured by the correlated error structure of the purchase utilities, i.e. choice in one category alters the utility of choices in other categories. Thus, if the covariance of the errors is positive, then an increase in the purchase utility of category  $i$  will lead to an increase in the purchase utility of category  $j$ . In other words, the error correlations capture the linkages between the uncontrollable factors that

drive joint purchases (Manchanda, Ansari, and Gupta, 1999). The unobserved factors that affect total expenditures of shopping trips are  $\varepsilon_{1ht} \sim \text{MVN}[0, \Sigma_{1(j \times j)}]$ . The unobserved incidence and expenditure errors are correlated, i.e.  $\text{cov}[\varepsilon_{1ht}, \varepsilon_{2ht}] = \Sigma_{12(j \times j)}$ , as will be explained in more detail below.

As explanatory variables, we only consider the own effects. Own effects show the impact of explanatory variables on the same category purchase. We do not consider the cross effects which reflect the change in purchase utility of category  $j$  due to the marketing actions of other categories, since these effects are likely to be small for the categories that we study. In addition; including those cross-effects would render our model very highly parameterized, where as our main interest is in representing the covariance of incidence and expenditure of multiple categories (Note that Manchanda, Ansari, and Gupta (1999) did not find any significant cross-effects of the incidence of four categories, except for detergent-softener and mix-cake category pairs, which are obviously related).

### **4.3.3 Individual Level Heterogeneity**

We include individual level heterogeneity into the model by considering household specific coefficients.

$$\beta_{1h} = Z_h \Theta_1 + \zeta_{1h}, \quad h = 1, \dots, H \quad (4.5a)$$

$$\beta_{2h} = Z_h \Theta_2 + \zeta_{2h}, \quad h = 1, \dots, H \quad (4.5b)$$

### Promotion Customization across Multiple Categories

We can write individual level parameters  $\beta_{1h}$  and  $\beta_{2h}$  as

$$\beta_{1h} \sim \text{MVN}(Z_h \Theta_1, \Lambda_1), \text{ and } \beta_{2h} \sim \text{MVN}(Z_h \Theta_2, \Lambda_2)$$

and  $m$  is the number of household-level explanatory variables. These variables are measured at the individual level and characterize a household's shopping behavior (see Rossi and Ainslie, 1998). While  $\Theta_1$  and  $\Theta_2$  indicate the impact of household level explanatory variables,  $\Lambda_1$  and  $\Lambda_2$  represent the unobserved sources of heterogeneity across households. In our application, we do not have any individual level explanatory variables, so  $Z_h$  contains an intercept term only. Since  $Z_h$  is equal to 1 for all  $h$ ,  $\Theta$  is the mean vector for the explanatory variables. The error terms are distributed as  $\zeta_{1h} \sim \text{MVN}(0, \Lambda_1)$ ,  $\zeta_{2h} \sim \text{MVN}(0, \Lambda_2)$ .

#### 4.3.4 Joint Model (Hierarchical Multivariate Type-2 Tobit Model)

The full model can now be represented as

$$Y_{ht}^* = \begin{bmatrix} Y_{1ht}^* \\ Y_{2ht}^* \end{bmatrix} = \begin{bmatrix} X_{1ht} & 0 \\ 0 & X_{2ht} \end{bmatrix} \begin{bmatrix} \beta_{1h} \\ \beta_{2h} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1ht} \\ \varepsilon_{2ht} \end{bmatrix} \quad (4.6)$$

$$\varepsilon_{ht} = \begin{bmatrix} \varepsilon_{1ht} \\ \varepsilon_{2ht} \end{bmatrix} \sim \text{MVN}\left(0, \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{bmatrix}\right) \quad (4.7)$$

$$\beta_h = \begin{pmatrix} \beta_{1h} \\ \beta_{2h} \end{pmatrix} \sim \text{MVN}(V_h \Theta, \Lambda) \quad (4.8)$$

$$\Lambda = \begin{bmatrix} \Lambda_1 & \text{Cov}(\Lambda_1, \Lambda_2) \\ \text{Cov}(\Lambda_2, \Lambda_1) & \Lambda_2 \end{bmatrix}, \quad \Theta = \begin{bmatrix} \Theta_1 & 0 \\ 0 & \Theta_2 \end{bmatrix}, \quad V_h = I_{(2 \times 2)} \otimes Z_h \quad (4.9)$$

#### **4.4 Estimation with MCMC**

We use Markov Chain Monte Carlo Methods (MCMC) to estimate our model. Chib (1992) describes MCMC methods for the standard tobit censored regression model, and proposes data-augmentation algorithms to simulate the unobserved variables  $Y_{ht}^*$  in every step of the Gibbs sampler. We apply those procedures to estimate our model. Details of an MCMC algorithm for a hierarchical multivariate type-2 tobit model are provided by Fox, Montgomery and Lodish (2004). For reasons of identifiability, in the multivariate probit model diagonal elements of the covariance matrix are set to ones, i.e.  $\sigma_{2ii}=1$ , for all  $i$ , so that the matrix  $\Sigma_2$  is a correlation matrix. The full conditional distribution of the correlation matrix is not tractable form and therefore not easy to draw. Liechty, Ramaswamy and Cohen (2000) draw R using griddy Gibbs sampler methods, and Barnard, McCulloch and Meng (2000) generate R using variants of the Metropolis-hasting algorithm. However these methods are computationally intensive. We estimate covariance matrix  $\Sigma_0$  ( $\Sigma_{01}$ ,  $\Sigma_{02}$  and  $\Sigma_{012}$ ) and coefficients  $\beta_h$  and post-process the draws of  $\Sigma$ ,  $\beta_h$  and  $\Theta$  with the use of a diagonal matrix C ( $C = \text{diag}(\Sigma_{02})^{-1/2}$ ) and obtained the correlation matrix  $\Sigma = C\Sigma_0C'$  (Edwards and Allenby, 2003). In the MCMC algorithm we use 50000 iterations and burn-in 20000 i.e. saved every fifth draws after that for real data set, and monitor convergence of the chain through plots of the hyper-parameters.

#### 4.4.1 Gibbs Sampling

Technical details of MCMC estimation of the hierarchical multivariate type-2 tobit model are similar to Fox et al. (2004), with the exception that the expenditures and purchase incidence of categories are dependent in our model. Moreover, individual level heterogeneity is considered not only through an intercept term ( $\beta_0$ ) as in Fox et al., but the effects of marketing actions (price and promotion coefficients,  $\beta_{1h_j}$  and  $\beta_{2h_j}$  respectively) are also allowed to be heterogeneous.  $Y_{h(j)t}$  illustrates expenditures or utilities of category  $i$ , and  $Y_{h(-j)t}$  is all the others. We need to specify conditional distributions of the relevant variables for Gibbs sampling. Natural conjugate priors are chosen for estimation. The stages in the Gibbs sampler are represented by  $s$  below, and the conditional draws are shown:

$$\Sigma^{(s+1)} \mid Y_{hjt}^{*(s)}, \beta_h^{(s)}, \Theta^{(s)}, \Lambda^{(s)}$$

$$Y_{1h(j)t}^{*(s+1)} \mid Y_{1h(-j)t}^{*(s)}, \beta_h^{(s)}, \Theta^{(s)}, \Lambda^{(s)}, \Sigma^{(s+1)}$$

$$Y_{2h(j)t}^{*(s+1)} \mid Y_{2h(-j)t}^{*(s)}, \beta_h^{(s)}, \Theta^{(s)}, \Lambda^{(s)}, \Sigma^{(s+1)}$$

$$\beta_h^{(s+1)} \mid Y_{ht}^{*(s+1)}, \Sigma^{(s+1)}, \Theta^{(s)}, \Lambda^{(s)}$$

$$\Theta^{(s+1)} \mid Y_{ht}^{*(s+1)}, \beta_h^{(s+1)}, \Sigma^{(s+1)}, \Lambda^{(s)}$$

$$\Lambda^{(s+1)} \mid Y_{ht}^{*(s+1)}, \beta_h^{(s+1)}, \Sigma^{(s+1)}, \Theta^{(s+1)}$$

4.4.1.1 Prior Distributions:

In Gibbs sampling, we need to specify prior distributions for the parameters of interest. We used noninformative priors.

1. The prior distribution of  $\Sigma^{-1}$  is Wishart  $W[\rho_1, R_0]$ , where  $\rho_1 = J * 2 + 2$ ,  $J$  is the number of categories, and  $R_0 = I$ .
2. The prior distribution of  $\Lambda^{-1}$  is  $W_p[\rho_2, R_1]$ , where  $\rho_2 = p + 2$ ,  $R_1 = I$ , and  $p$  is the rank of  $\Theta$ .
3. The prior distribution of  $\Theta$  is a multivariate normal  $MVN[\Theta_0, V_\Theta]$  where  $\Theta_0 = 0$  and  $V_\Theta = \text{diag}(10^{-3})$ .

4.4.1.2 Full Conditional Distributions:

1. The full conditional distribution of the residual covariance matrix,  $\Sigma^{(s+1)}$  is inverse wishart,  $\Sigma^{-1(s+1)} \sim W[N + \rho_1, S]$ , where  $N$  is the total number of observations ( $N = H * T$ ) and  $S = (Y_h^* - X_h \beta_h^{(s)})'(Y_h^* - X_h \beta_h^{(s)}) + R_0$ .

2. The full conditional distribution of latent utilities ( $Y_2$ ) for the probit part of the model is a truncated multivariate normal,  $Y_{2\text{hit}}^{*(s+1)} \sim \text{TMVN}(\mu_{20}, \Sigma_{20})$ , with mean  $\mu_{20}$  and variance  $\Sigma_{20}$ , which are shown below. If the indicator variable  $Y_2 = 1$ , then  $Y_2^*$  is drawn from a normal distribution, truncated below at 0. Otherwise,  $Y_2^*$  is drawn from a normal distribution, truncated above at 0. We have:

*Promotion Customization across Multiple Categories*

$$Y_{ht}^* = \begin{bmatrix} Y_{hjt}^* \\ \dots \\ Y_{h,(-),t}^* \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_{jj} & | & \sigma_{j(-)} \\ \dots & + & \dots \\ \sigma_{(-)j} & | & \Sigma_{(-)j(-)} \end{bmatrix}, \text{ and}$$

$$\mu_{20} = X_{2hjt} \beta_{2h}^{*(s)} + (Y_{2h,(-),t}^{*(s)} - E(Y_{2h,(-),t}^{*(s)})) \Sigma_{2(-)j(-)}^{-1(s)} \sigma_{2j(-)}^{(s)}$$

$$\Sigma_{20}^{(s+1)} = \sigma_{2jj}^{(s+1)} - \sigma_{2j(-)}^{(s+1)} \Sigma_{2(-)j(-)}^{-1(s+1)} \sigma_{2(-)j}^{(s+1)}$$

3.  $Y_1^*$  is  $Y_1$  if  $y_1 > 0$ , otherwise  $Y_1^*$  is drawn from a normal distribution, truncated above at 0. The full conditional distribution of expenditures is a truncated multivariate normal distribution,  $Y_{1hjt}^{*(s+1)} \sim \text{TMVN}(\mu_{10}, \Sigma_{10})$ , with mean  $\mu_{10}$  and variance  $\Sigma_{10}$ , where

$$\mu_{10}^* = X_{1hjt} \beta_{1h}^{*(s)} + (Y_{1h,(-),t}^{*(s)} - E(Y_{1h,(-),t}^{*(s)})) \Sigma_{1(-)j(-)}^{-1(s)} \sigma_{1j(-)}^{(s)}$$

$$\Sigma_{10}^{(s+1)} = \sigma_{1jj}^{(s+1)} - \sigma_{1j(-)}^{(s+1)} \Sigma_{1(-)j(-)}^{-1(s+1)} \sigma_{1(-)j}^{(s+1)}$$

The inverse cdf method is used to draw truncated normal values for  $Y_1 > 0$ . It is explained below (see also Fox et al., 2004):

- Compute the upper limit for the uniform interval,  
 $L = \Phi\left[\frac{0 - E[Y_{1hjt}^* | Y_{1h,(-),t}^*, \beta_h, \Theta, \Sigma, \Lambda]}{(\sigma_{jj} - \sigma_{j(-)} \Sigma_{(-)j(-)}^{-1} \sigma_{j(-)})}\right]$ , where  $\Phi[.]$  represents the Normal cumulative distribution function (cdf) and  
 $E[Y_{1hjt}^* | Y_{1h,(-),t}^*, \beta_h, \Theta, \Sigma, \Lambda] = X_{1hjt} \beta_h^{*(s)} + (Y_{1h,(-),t}^{*(s)} - E(Y_{1h,(-),t}^{*(s)})) \Sigma_{(-)j(-)}^{-1(s)} \sigma_{j(-)}^{(s)}$
- Draw a uniform variate,  $U \sim \text{Uniform}(0, L)$
- Compute the value of the uniform draw:  
 $Y_{1i}^* = \Phi^{-1}(U)(\sigma_{jj} - \sigma_{j(-)} \Sigma_{(-)j(-)}^{-1} \sigma_{j(-)}) + E[Y_{1i}^* | Y_{1(-),t}^*, \beta, \Theta, \Sigma, \Lambda]$

4. The full conditional distribution of individual level coefficients  $\beta_h$  is multivariate normal,  $\beta_h^{(s+1)} \sim \text{MVN}[M_c, V_c]$ , where

$$M_c = V_c \text{vec}((X_h' Y_h) \times S + \Sigma^{-1(s+1)} (Z\Theta^{(s)})'), \text{ and } V_c = (S^{-1} \otimes (X_h' X_h) + \Sigma^{-1(s+1)})^{-1}.$$

5. The full conditional distribution of  $\Theta$  is a multivariate normal,  $\Theta^{(s+1)} \sim \text{MVN}[M_c, V_c]$  with mean  $M_c = \text{vec}(Z_h' \beta_h^{(s+1)} \times \Lambda^{-1(s)}) \times V_c$  and variance  $V_c = (\Lambda^{-1(s)} \otimes (Z_h' Z_h) + V_\Theta)^{-1}$ .

6. The full conditional distribution of error of individual coefficients,  $\Lambda$  is inverse wishart,  $\Lambda^{-1(s+1)} \sim W(H + \rho_2, S)$ , where  $H$  is the number of subjects and  $S = (\beta_h^{(s+1)} - Z\Theta^{(s+1)})'(\beta_h^{(s+1)} - Z\Theta^{(s+1)}) + R_1$ .

## **4.5 Customized Promotions Design**

The typical retailer's decision problem is to choose the right categories to promote, since not all of them can be promoted, or necessarily should be promoted, at the same time. This is in contrast with the manufacturers' planning problem, in which each product line and brand has its own promotional plan. The retailer needs certain selection criteria (common ones are store traffic generation, profitability of the item, revenue it generates or the image it creates). The components of customized promotion decisions are when, where, what, and to whom to promote. In this section, we focus on the optimization of promotion decisions regarding which product categories from many possible to promote to whom, in the

### *Promotion Customization across Multiple Categories*

online shopping venue. We maximize the retailer's revenue, based on the model of the two related consumer decisions –in which categories to purchase and how much to spend in each of them. We will use combinatorial optimization routines to customize the promotion plan to individual customers.

We assume that the categories that we choose offer a price promotion (the proposed approach can also be applied at the point of purchase for 'in store' targeted coupon distribution). Since we have many categories and the unit measure of size for these categories differs between categories, we consider consumer expenditure as a criterion function to optimize. Cost and margin data are seldom available, and are not available to us in this study. In marketing, revenue has been used as a criterion for establishing pricing policies (see Anjos et al., 2005), and has been quite popular, for example, in the airline industry. As Rossi et al. (1996) state, "Any successful customization approach must deal directly with the problem of partial information and take parameter uncertainty into account in the decision problem." Therefore, we use a Bayesian decision theoretic approach (Dorfman, 1997) to our promotion allocation problem. That is, we estimate our objective function --retailer revenue-- in every draw of the Gibbs sampler, which enables us to minimize expected loss or, equivalently, maximize expected revenue, across the draws integrating out parameter uncertainty. To compute the optimization criterion, we estimate the a-posteriori expected difference in category expenditure for each consumer if we promote, respectively if we do not promote, the category in question. After we estimate the expected expenditures in these two cases and compute the difference for each category at every MCMC draw, we

generate promotion designs by choosing that allocation of promotions that maximizes, for each individual, the revenue difference arising from promoting across categories. In doing so, we reflect restrictions on the number of categories to promote, say  $P$ , that operate in most applied situations. The question then becomes which  $P$  categories to choose for each consumer. For each consumer, we estimate the expected difference in expenditure if we promote a category and do not promote it, in order to generate all possible optimal promotion designs for each consumer, and then accept the design that has the maximum total expenditure. This procedure will give us the information of which categories to promote to whom.

In generating promotion designs, we set restrictions on the total number of categories to promote: at most  $P$  categories will be promoted at the same time, since obviously the online retailer cannot promote all categories at the same time. This approach is needed because of the space limitations on web pages, cash register receipts, and e-mail messages for displaying electronic coupons to customers. Additionally, these restrictions reduce the number of possible promotion plans (contained in the candidate promotion-plan matrix,  $C$ ), which reduces computing time to find an optimal design<sup>12</sup>. In our optimization problem, instead of a candidate set of a size  $2^J$ , we will

---

<sup>12</sup> For example, the candidate set ( $C$ ) for  $J=4$  categories (columns) with  $P=3$  being chosen to promote consists of 4 candidate promotion plans (rows):

$$C = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

*Promotion Customization across Multiple Categories*

use a candidate set (C), with size  $\binom{J}{P}$ . For example, in our application, there are 4368 possible promotion plans, choosing five categories to promote out of sixteen. We define the criterion function as a function of expected consumer expenditures, which are a function of the price and promotion variables. Assume that all parameters of the model are collected in  $\psi$ . Our optimization problem for each consumer  $h$  is defined as:

$$\operatorname{argmax}_{\{d_{hj}\}} E[\pi_i] = E_{\psi} \left[ \sum_{j=1}^J \left( E[Y_{1hj}^1] - E[Y_{1hj}^0] \right)^{d_{hj}} \right] \quad (4.10)$$

$$\text{s.t.} \quad \sum_{j=1}^J d_{hj} = K ; j=1, \dots, J \quad h=1, \dots, H, \text{ with,}$$

$$d_{hj} = \begin{cases} 1 & \text{if category } j \text{ is chosen} \\ 0 & \text{if category } j \text{ is not chosen} \end{cases}$$

$d_{hj}$  is determined by a search over all possible designs in the candidate matrix C, using the modified Federov algorithm as illustrated below.  $E[Y_{1hj}^1]$  is the expected expenditure of a customer  $h$  given that we promote  $j$ , and  $E[Y_{1hj}^0]$  is the expected expenditure of a customer  $h$  given that we do not promote  $j$ . We estimate these quantities for each consumer and in each Gibbs iterations. So at draw  $s$  of the Gibbs sampler, we obtain  $\psi^{(s)}$ , which enables us to compute  $E^{(s)}[Y_{1hj}^0 | X'_{hj}, \psi^{(s)}]$  and  $E^{(s)}[Y_{1hj}^1 | X''_{hj}, \psi^{(s)}]$ . Here  $X'_{hj}$  and  $X''_{hj}$  are our design matrices, in which the promotion variables are set to 0 and 1, respectively, for category  $j$  and consumer  $h$ . We estimate the

expected expenditure of category j if we promote it, through the expectation of  $Y_{1hj}$  conditional on  $X'_{hj}$  :

$$E[Y_{1hj}^0 | X'_{hj}] = P(Y_{2hj}^* > 0) \cdot E(Y_{1hj} | X'_{hj}, Y_{2hj}^* > 0) \quad (4.11a)$$

$$= \left( \frac{1}{\sigma_{2jj}} \right) \Phi \left( \frac{Y_{2hj} - X'_{hj} \beta_{2h}}{\sigma_{2jj}} \right) \times \left( X'_{hj} \beta_{1h} + \frac{\Sigma_{2j(-j)}}{\sqrt{\Sigma_{2jj} \Sigma_{2(-j)(-j)}}} \sigma_{2jj} \frac{\phi(-X'_{hj} \beta_{2h})}{1 - \Phi(-X'_{hj} \beta_{2h})} \right) \quad (4.11b)$$

Note that we use a correlation matrix in the MCMC estimation for identification reasons and the diagonals elements are equal to 1 ( $\sigma_{1jj}=1$ ,  $\sigma_{2jj}=1, \dots$ ). A similar expression is obtained for  $E^{(s)}[Y_{1hj}^1 | X''_{hj}, \psi^{(s)}]$ . Here we have used:

$$\Sigma_1 = \begin{bmatrix} \sigma_{1jj} & | & \sigma_{1j(-j)} \\ \text{---} & | & \text{---} \\ \sigma_{1(-j)j} & | & \Sigma_{1(-j)(-j)} \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} \sigma_{2jj} & | & \sigma_{2j(-j)} \\ \text{---} & | & \text{---} \\ \sigma_{2(-j)j} & | & \Sigma_{2(-j)(-j)} \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{jj} & | & \Sigma_{j(-j)} \\ \text{---} & | & \text{---} \\ \Sigma_{(-j)j} & | & \Sigma_{(-j)(-j)} \end{bmatrix}.$$

Now we need to take the expectation of (4.11b) over the distribution of the parameters. The expectation for, for example,  $Y_{1hj}^0$  equals

$$E_{\psi} [E[Y_{1hj}^0 | X'_{hj}]] = \int E[Y_{1hj}^0 | X'_{hj}, \psi] f(\psi) d\psi.$$

This enables us to compute the two components of the criterion function across the iterates of the Gibbs sampler as:

$$E_{\psi} [E[Y_{1hj}^1 | X''_{hj}]] \approx \sum_s \frac{E[Y_{1hj}^1 | X''_{hj}, \psi^{(r)}]}{S}$$

$$E_{\psi} [E[Y_{1hj}^0 | X''_{hj}]] \approx \sum_s \frac{E[Y_{1hj}^0 | X'_{hj}, \psi^{(r)}]}{S} \quad (4.12)$$

#### **4.5.1 Modified Fedorov Algorithm**

Exchange algorithms are mainly used for finding plans in experimental designs where all the decision variables can be set at specified values in combinations determined by the plan. In our decision, our variable is 1/0 variable of whether or not to promote a category out of J categories. There are  $2^{16}-1$  (=65535) different promotion plans for sixteen categories. Since our promotion plans have the restrictions that only five categories be promoted, the candidate promotion plan matrix is reduced to 4368 different promotion plans. Now we want to choose the best plan for each household from this candidate set. The Federov (1972) algorithm is used for this purpose, and is described below in the context of the specific design criterion chosen:

1. A candidate list of all feasible combinations of the promotion plans is constructed.
2. One combination is randomly selected from the candidate list for each household as a starting plan. We calculate the value of design criterion ( $E(\pi_i)$ ) for this plan.
3. We exchange each combination of promotion plans for each household with the remaining promotion plans in the candidate list, and calculate the value of the design criterion for the new plan. This process is repeated, and the exchange that leads to the largest reduction in the design criterion is accepted. In the modified Federov algorithm (Cook and

Nachtsheim, 1980), any exchange that reduces the value of design criterion is made as soon as it is found, which speeds up the algorithm.

4. Until the improvement in design criterion is smaller than some specified tolerance, we repeat step 3.

5. We repeat steps 2 to 4 four times to try to avoid local optima, and the best plan found is used as an optimal promotion design.

## 4.6 Synthetic Data Results for Model Estimation

In this section, we discuss the results of three simulation studies investigating the performance of the models and estimation algorithms. In these studies, we checked the performance of models and algorithms in three cases: the simulation results for the simple multivariate type-2 tobit model when there is no covariance between incidence and expenditure, and without individual heterogeneity; with covariance and no individual heterogeneity; with covariance and individual heterogeneity. The model is estimated with the Gibbs sampling for two categories, and 100 households in three cases. Synthetic data are generated with known parameters (true values), which were compared to the estimated parameter values. That is, the design of the simulation study mimics the structure of the empirical data, so that good recovery in the simulation gives us confidence for the performance of the algorithm with the real data.

### 4.6.1 No Covariance between Incidence and Expenditure, No Individual Heterogeneity

In the first case, we assume that all subjects have the same preference ( $\beta$ ) coefficients, i.e. no individual heterogeneity, and that there is no interdependence between purchase incidence and expenditure between categories. We assume that there are two explanatory variables. The error matrix ( $\Sigma$ ) is in correlation form, and for this reason, incidence diagonal  $\sigma$  values are equal to 1. As we can see, parameter estimates of correlation (Table 4.1),  $\beta$ 's (Table 4.2) and expenditures'  $\sigma$  (Table 4.3) are close to true ones, and results reveal a satisfactory performance of the Gibbs sampling algorithm.

Table 4.1: Simulation results for the correlation ( $\Sigma$ ) matrix

category	True $\Sigma$				Posterior Mean $\Sigma$				Posterior SE $\Sigma$			
	expenditure(I)		incidence(II)		expenditure(I)		incidence(II)		expenditure(I)		incidence(II)	
	1	2	1	2	1	2	1	2	1	2	1	2
(I) 1	1.00	-0.29	0.00	0.00	1.00	-0.10	-0.02	0.04	0.00	0.03	0.05	0.04
(I) 2	-0.29	1.00	0.00	0.00	-0.10	1.00	-0.04	0.02	0.03	0.00	0.05	0.05
(II) 1	0.00	0.00	1.00	0.30	-0.02	-0.04	1.00	0.34	0.05	0.05	0.00	0.04
(II) 2	0.00	0.00	0.30	1.00	0.04	0.02	0.34	1.00	0.04	0.05	0.04	0.00

Table 4.2: Simulation results for  $\beta_i$ 's

<u>Expenditure</u>				<u>Incidence</u>			
True $\beta_{1i}$				True $\beta_{2i}$			
-0.508	1.949	0.259	-0.679	0.049	-0.148	-0.183	0.042
Posterior $\beta_{1i}$				Posterior $\beta_{2i}$			
-0.700	2.138	0.220	-0.644	-0.024	-0.138	-0.139	0.055
Posterior SE of $\beta_{1i}$				Posterior SE of $\beta_{2i}$			
0.212	0.200	0.126	0.127	0.192	0.192	0.203	0.203

Table 4.3: Simulation results for expenditure's standard deviation ( $\sigma_{1ij}$ )

	True	Posterior Mean	Posterior SE
	<b>expenditure</b>	5.766	5.721
	3.873	3.667	0.092

#### 4.6.2 With Covariance between Incidence and Expenditure, No Individual Heterogeneity

In this case, we still assume that all subjects have the same preference ( $\beta$ ) coefficients, i.e. no individual heterogeneity, and we have two explanatory variables, but there is interdependence between purchase incidence and expenditure between categories (see correlation between expenditure and incidence in the correlation matrix, Table 4.4). We obtain satisfactory results for parameter estimates ( $\Sigma$  in Table 4.4,  $\beta$ 's in Table 4.5 and expenditures'  $\sigma$  in Table 4.3), i.e. the posterior means are close to true parameter values.

Table 4.4: Simulation results for the correlation ( $\Sigma$ ) matrix

category	True $\Sigma$				Posterior Mean $\Sigma$				Posterior SE $\Sigma$			
	expenditure(I)		incidence(II)		expenditure(I)		incidence(II)		expenditure(I)		incidence(II)	
	1	2	1	2	1	2	1	2	1	2	1	2
(I) 1	1.00	-0.29	0.22	0.99	1.00	-0.31	0.26	0.99	0.00	0.03	0.04	0.00
(I) 2	-0.29	1.00	0.31	-0.32	-0.31	1.00	0.29	-0.29	0.03	0.00	0.04	0.03
(II) 1	0.22	0.31	1.00	0.30	0.26	0.29	1.00	0.31	0.04	0.04	0.00	0.04
(II) 2	0.99	-0.32	0.30	1.00	0.99	-0.29	0.31	1.00	0.00	0.03	0.04	0.00

Table 4.5: Simulation results for  $\beta_i$ 's

<u>Expenditure</u>				<u>Incidence</u>			
True $\beta_{1i}$				True $\beta_{2i}$			
-1.228	0.497	0.437	1.830	0.243	0.453	0.064	0.270
Posterior $\beta_{1i}$				Posterior $\beta_{2i}$			
-1.052	0.638	0.373	1.820	0.240	0.494	0.069	0.263
Posterior SE of $\beta_{1i}$				Posterior SE of $\beta_{2i}$			
0.156	0.141	0.128	0.116	0.173	0.198	0.177	0.157

Table 4.6: Simulation results for expenditure's standard deviation ( $\sigma_{1ii}$ )

	Posterior		
	True	Mean	SE
<b>expenditure</b>	5.766	5.908	0.123
	3.873	3.885	0.093

### 4.6.3 With Covariance between Incidence and Expenditure, and Individual Heterogeneity

In the final case, we use two explanatory variables, but there is interdependence between purchase incidence and expenditure between categories, and all subjects have unique preference ( $\beta_n$ ) coefficients, i.e. the model now accommodates individual heterogeneity. We do not report individual  $\beta$ 's for convenience, since we have 100 subjects. We report the correlation matrix of the errors for expenditure and incidence ( $\Sigma$ ), the mean level preference coefficients ( $\Theta$ ), the standard deviation of expenditure errors ( $\sigma$ ), and the covariance ( $\Lambda$ ) of the marketing mix variables. We again obtain satisfactory results for all parameter estimates. Estimates for  $\Sigma$  are

presented in Table 4.7, for  $\Theta$  in Table 4.8, for  $\sigma$  in Table 4.9, and for  $\Lambda$  in Table 4.10. Note that  $\Lambda$  values for the incidence part are not as good as  $\Lambda$  values for the expenditure part in Table 4.10. The simulation results are satisfactory, and now we can apply this method to the real data set with sixteen product categories.

Table 4.7: Simulation results for the correlation ( $\Sigma$ ) matrix

category	True $\Sigma$				Posterior Mean $\Sigma$				Posterior SE $\Sigma$			
	expenditure(I)		incidence(II)		expenditure(I)		incidence(II)		expenditure(I)		incidence(II)	
	1	2	1	2	1	2	1	2	1	2	1	2
(I) 1	1.00	-0.29	0.22	0.99	1.00	-0.31	0.22	0.98	0.00	0.04	0.04	0.00
(I) 2	-0.29	1.00	0.31	-0.32	-0.31	1.00	0.34	-0.26	0.04	0.00	0.04	0.04
(II) 1	0.22	0.31	1.00	0.30	0.22	0.34	1.00	0.34	0.04	0.04	0.00	0.04
(II) 2	0.99	-0.32	0.30	1.00	0.98	-0.26	0.34	1.00	0.00	0.04	0.04	0.00

Table 4.8: Simulation results for  $\Theta$

<u>Expenditure</u>				<u>Incidence</u>			
True $\beta_{1i}$				True $\beta_{2i}$			
0.182	-1.801	-1.086	1.434	-0.227	-0.104	0.308	-0.389
Posterior $\beta_{1i}$				Posterior $\beta_{2i}$			
0.134	-1.612	-1.381	1.414	-0.237	-0.159	0.267	-0.406
Posterior SE of $\beta_{1i}$				Posterior SE of $\beta_{2i}$			
0.212	0.196	0.179	0.195	0.377	0.383	0.367	0.400

Promotion Customization across Multiple Categories

Table 4.9: Simulation results for expenditure's standard deviation ( $\sigma_{1ij}$ )

	True	Posterior Mean	Posterior SE
<b>expenditure</b>	5.766	5.829	0.136
	3.873	3.836	0.094

Table 4.10: Simulation results for  $\Lambda$

True $\Lambda$							
1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	1
Posterior $\Lambda$							
0.827	-0.226	0.132	0.120	0.192	0.036	0.744	-0.312
-0.226	0.968	-0.052	-0.253	-0.492	0.024	-0.204	0.960
0.132	-0.052	1.082	-0.076	-0.168	-0.048	0.084	-0.228
0.120	-0.253	-0.076	1.278	-0.012	0.432	-0.060	-0.504
0.192	-0.492	-0.168	-0.012	1.342	0.024	0.180	-0.132
0.036	0.024	-0.048	0.432	0.024	1.262	-0.012	0.132
0.744	-0.204	0.084	-0.060	0.180	-0.012	1.108	-0.085
-0.312	0.960	-0.228	-0.504	-0.132	0.132	-0.085	1.180
SE of Posterior $\Lambda$							
0.417	0.263	0.241	0.241	0.344	0.337	0.339	0.344
0.263	0.339	0.216	0.192	0.325	0.317	0.324	0.323
0.241	0.216	0.275	0.169	0.275	0.277	0.295	0.282
0.241	0.192	0.169	0.219	0.258	0.261	0.268	0.266
0.344	0.325	0.275	0.258	0.770	0.554	0.593	0.575
0.337	0.317	0.277	0.261	0.554	0.789	0.579	0.570
0.339	0.324	0.295	0.268	0.593	0.579	0.909	0.621
0.344	0.323	0.282	0.266	0.575	0.570	0.621	0.897

## 4.7 Data Description

We have purchase data from a random sample of customers of a leading online grocery retailer. The data are from May 1996 to July 1997, from a total of 279 households. We have the purchase history data of 16 product-categories over time, and the incidence frequencies are shown in Table 4.11. We know that 133 households shopped regularly from one of these product-categories from 1996-1997. We have a total of 4281 shopping trips, and at least one category was purchased on 3632 occasions during these trips. We use 62 shopping weeks. The numbers of pair-wise purchases for all 16 product-category pairs are shown in Table 4.12. Detergents, paper towels and toilet paper are the most frequently purchased categories and also the most promoted categories, as is evident from Table 4.11. Squeeze margarine, butter, allergy medicines and coffee instant decaf are the most infrequently purchased product categories (Table 4.11). The online purchase data contains the number of units of SKUs bought from each of the 16 categories, size, price paid per unit, and whether or not the price reflected a deal purchase. If the same SKU was bought more than once for the same price in one shopping trip, it was treated as a single purchase by aggregating the quantities. However, if different SKUs in the same product category were chosen on the same shopping trip, we randomly selected one of them. The number of shopping trips ranged from 9 to 69, with the average being 32 per household. The average total expenditure per shopping trip is \$125 per household. This

### *Promotion Customization across Multiple Categories*

calculation included the total expenditure of those consumers for whom an itemization of expenditure per product category was not available, i.e. purchases of “unknown categories.”

From the correlation matrix of purchase incidence of product categories (Table 4.13), paper towel tissue and toilet paper tissue has the highest correlation (0.413). We observe that paper towel tissue and paper toilet tissue (0.413), paper towel tissue and laundry detergent (0.262), toilet paper tissue and laundry detergent (0.243), spaghetti sauce and toilet paper tissue (0.226), toilet paper tissue and soft margarine (0.219), spaghetti sauce and soft margarine (0.205) and finally toilet paper tissue and soap (0.203) are purchased the most frequently together compared to other category pairs. Again in this table, there are quite a few negative but low correlation values. The noticeable negatively correlated category pairs are: Squeeze margarine and stick margarine are negatively correlated with butter (-0.010 and -0.022, they are 3 and 17 times purchased together, respectively). Note that if we estimate the correlation matrix with only purchase weeks or including nonpurchase weeks, we obtain almost the same results.

The expenditure correlation matrix is presented in Table 4.14. Again toilet paper tissue and paper towel tissue have the highest correlation (0.295), but this correlation is not as high as their incidence correlation (0.413). The other highly correlated category pairs are toilet paper tissue and laundry detergent (0.256), paper towel tissue and laundry detergent (0.251), soft margarine and crackers (0.224), paper towel tissue and butter (0.208), toilet paper tissue and stick margarine (0.157), soap and toilet

paper tissue (0.149), and, finally, spaghetti sauce and toilet paper tissue (0.148).

Table 4.15 illustrates the correlation of purchase incidence and expenditures of categories. The correlation of paper toilet tissue incidence and expenditure is equal to 0.787 and paper towel tissue is 0.750, which are the lowest. This means consumers purchase these two categories frequently but do not spend much. There may be two explanations for that: these categories are promoted a lot (which is actually confirmed by the descriptive statistics in Table 4.12), or consumers purchase these categories in high quantities.

In Table 4.16, we illustrate some descriptive statistics for purchase and nonpurchase weeks. We can easily see that except for allergy tablets and squeeze margarine (average the same prices in two cases), the prices of remaining categories are higher for nonpurchase weeks than the purchase weeks.

*Promotion Customization across Multiple Categories*

Table 4.11: Descriptive statistics of the online purchase

<b>Category</b>	<b># of SKU in Category</b>	<b>Purchased # of SKU</b>	<b># of Purchase on Promotion</b>	<b># of Category Purchase</b>	<b>% of Purchase on Promotion</b>
Allergy medicine	79	8	1	17	0.059
Butter	32	9	227	686	0.331
Coffee Gr. Decaf	85	24	11	159	0.069
Coffee Gr. Regular	187	40	27	224	0.121
Coffee Ins. Decaf	24	8	0	30	0.000
Coffee Ins. Regular	68	20	3	87	0.034
Cold medicine	222	27	6	48	0.125
Crackers	32	14	60	225	0.267
Laundry	179	48	195	833	0.234
Margarine Soft	62	24	124	638	0.194
Margarine Squeeze	4	2	36	60	0.600
Margarine Stick	42	15	97	358	0.271
Paper Toilet	114	38	240	1954	0.123
Paper Towel	74	28	246	1863	0.132
Soap	137	45	72	411	0.175
Spaghetti Sauce	249	76	193	744	0.259

Table 4.12: Descriptive statistics: joint purchase frequencies

Categories	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Allergy	17															
2 Butter	0	686														
3 Coffee GD.	0	24	159													
4 Coffee GR.	1	32	26	224												
5 Coffee ID.	0	4	1	4	30											
6 Coffee IR.	0	8	2	5	8	87										
7 Cold	4	3	3	1	0	2	48									
8 Cracker	0	49	15	15	0	6	1	225								
9 Laundry	4	145	20	52	10	21	13	32	833							
10 Marg. Soft	3	80	27	53	7	8	9	44	152	638						
11 Marg. Sqz.	2	3	1	1	0	0	0	2	8	9	60					
12 Marg. Stick	1	17	21	27	2	5	3	32	72	80	4	358				
13 Paper Toilet	11	365	78	102	16	39	27	109	444	348	27	171	1954			
14 Paper Towel	7	385	72	107	11	40	15	127	458	277	18	177	1035	1863		
15 Soap	4	69	16	30	1	10	5	36	98	72	6	67	245	213	411	
16 Spag. Sauce	2	127	41	58	4	13	7	61	143	186	11	79	394	331	103	744

Promotion Customization across Multiple Categories

Table 4.13: Descriptive statistics: bivariate correlations of purchase incidence

Category	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Allergy	1															
2 Butter	-0.014	1														
3 Coffee GD.	-0.006	<b>0.039</b>	1													
4 Coffee GR.	0.009	<b>0.048</b>	<b>0.118</b>	1												
5 Coffee ID.	-0.003	0.011	0.006	<b>0.039</b>	1											
6 Coffee IR.	-0.005	0.004	0.003	0.019	<b>0.151</b>	1										
7 Cold	<b>0.137</b>	-0.005	<b>0.024</b>	-0.003	-0.005	<b>0.023</b>	1									
8 Cracker	-0.008	<b>0.096</b>	<b>0.063</b>	<b>0.045</b>	-0.010	<b>0.026</b>	-0.003	1								
9 Laundry	0.020	<b>0.119</b>	0.012	<b>0.073</b>	<b>0.047</b>	<b>0.049</b>	<b>0.043</b>	<b>0.028</b>	1							
10 Marg. Soft	0.017	<b>0.051</b>	<b>0.049</b>	<b>0.100</b>	<b>0.035</b>	0.006	<b>0.032</b>	<b>0.074</b>	<b>0.135</b>	1						
11 Marg. Sqz.	<b>0.059</b>	-0.010	-0.002	-0.006	-0.005	-0.009	-0.007	0.003	0.009	<b>0.023</b>	1					
12 Marg. Stick	0.003	<b>-0.022</b>	<b>0.062</b>	<b>0.063</b>	0.007	0.007	0.007	<b>0.081</b>	<b>0.076</b>	<b>0.117</b>	0.010	1				
13 Paper Toilet	<b>0.044</b>	<b>0.219</b>	<b>0.086</b>	<b>0.094</b>	<b>0.043</b>	<b>0.052</b>	<b>0.059</b>	<b>0.111</b>	<b>0.243</b>	<b>0.219</b>	<b>0.044</b>	<b>0.129</b>	1			
14 Paper Towel	0.021	<b>0.250</b>	<b>0.079</b>	<b>0.105</b>	0.021	<b>0.059</b>	0.017	<b>0.142</b>	<b>0.262</b>	<b>0.152</b>	0.016	<b>0.143</b>	<b>0.413</b>	1		
15 Soap	<b>0.039</b>	<b>0.080</b>	<b>0.033</b>	<b>0.072</b>	-0.005	<b>0.031</b>	0.019	<b>0.092</b>	<b>0.109</b>	<b>0.087</b>	<b>0.020</b>	<b>0.138</b>	<b>0.203</b>	<b>0.164</b>	1	
16 Spag. Sauce	0.004	<b>0.107</b>	<b>0.086</b>	<b>0.106</b>	0.009	0.021	0.015	<b>0.111</b>	<b>0.100</b>	<b>0.205</b>	<b>0.028</b>	<b>0.100</b>	<b>0.226</b>	<b>0.173</b>	<b>0.129</b>	1

Bold categories are significantly correlated with 0.01. Bold and italic categories are significantly correlated with 0.05.

Table 4.14: Descriptive statistics: bivariate correlations of expenditure

Category	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Allergy	1															
2 Butter	-0.011	1														
3 Coffee GD.	-0.005	<b>0.027</b>	1													
4 Coffee GR.	0.005	<b>0.039</b>	<b>0.118</b>	1												
5 Coffee ID.	-0.002	-0.002	0.004	<b>0.037</b>	1											
6 Coffee IR.	-0.004	-0.004	0.001	0.010	<b>0.139</b>	1										
7 Cold	<b>0.077</b>	-0.004	0.008	-0.003	-0.004	<b>0.023</b>	1									
8 Cracker	-0.005	<b>0.057</b>	<b>0.043</b>	<b>0.031</b>	-0.008	<b>0.025</b>	-0.004	1								
9 Laundry	<b>0.026</b>	<b>0.118</b>	<b>0.027</b>	<b>0.053</b>	<b>0.029</b>	<b>0.058</b>	<b>0.030</b>	0.017	1							
10 Marg. Soft	0.002	<b>0.040</b>	<b>0.050</b>	<b>0.086</b>	<b>0.046</b>	0.004	<b>0.046</b>	<b>0.224</b>	<b>0.094</b>	1						
11 Marg. Sqz.	<b>0.056</b>	-0.003	-0.006	-0.006	-0.005	-0.007	-0.006	0.000	0.017	0.014	1					
12 Marg. Stick	<b>0.082</b>	<b>-0.024</b>	<b>0.050</b>	<b>0.046</b>	0.001	-0.001	0.000	<b>0.049</b>	<b>0.129</b>	<b>0.064</b>	0.001	1				
13 Paper Toilet	0.010	<b>0.142</b>	<b>0.050</b>	<b>0.090</b>	<b>0.046</b>	<b>0.094</b>	<b>0.040</b>	<b>0.085</b>	<b>0.256</b>	<b>0.208</b>	<b>0.034</b>	<b>0.143</b>	1			
14 Paper Towel	0.011	<b>0.208</b>	<b>0.079</b>	<b>0.063</b>	0.014	<b>0.070</b>	0.000	<b>0.100</b>	<b>0.251</b>	<b>0.100</b>	<b>0.005</b>	<b>0.157</b>	<b>0.295</b>	1		
15 Soap	0.016	<b>0.059</b>	<b>0.028</b>	<b>0.043</b>	0.003	<b>0.052</b>	0.000	<b>0.065</b>	<b>0.099</b>	<b>0.070</b>	<b>0.017</b>	<b>0.141</b>	<b>0.149</b>	<b>0.089</b>	1	
16 Spag. Sauce	-0.001	<b>0.070</b>	<b>0.087</b>	<b>0.068</b>	-0.001	<b>0.048</b>	0.008	<b>0.103</b>	<b>0.062</b>	<b>0.159</b>	<b>0.017</b>	<b>0.052</b>	<b>0.148</b>	<b>0.104</b>	<b>0.092</b>	1

Bold categories are significantly correlated with 0.01. Bold and italic categories are significantly correlated with 0.05.

Promotion Customization across Multiple Categories

4.15: Descriptive statistics: bivariate correlations of expenditure&purchase incidence

Expenditure	Purchase Incidence															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Allergy	<b>0.869</b>	-0.012	-0.006	0.003	-0.002	-0.004	<b>0.085</b>	-0.007	0.009	0.004	<b>0.065</b>	<b>0.024</b>	<b>0.040</b>	0.021	<b>0.025</b>	0.001
2 Butter	-0.013	<b>0.924</b>	<b>0.027</b>	<b>0.039</b>	0.005	0.003	-0.007	<b>0.077</b>	<b>0.117</b>	<b>0.046</b>	-0.005	-0.014	<b>0.200</b>	<b>0.229</b>	<b>0.065</b>	<b>0.093</b>
3 Coffee GD.	-0.006	<b>0.041</b>	<b>0.956</b>	<b>0.123</b>	0.008	0.000	0.013	<b>0.058</b>	0.015	<b>0.049</b>	-0.005	<b>0.047</b>	<b>0.078</b>	<b>0.078</b>	<b>0.036</b>	<b>0.089</b>
4 Coffee GR.	0.012	<b>0.049</b>	<b>0.105</b>	<b>0.941</b>	<b>0.046</b>	0.013	0.001	<b>0.040</b>	<b>0.070</b>	<b>0.100</b>	-0.006	<b>0.068</b>	<b>0.093</b>	<b>0.095</b>	<b>0.070</b>	<b>0.109</b>
5 Coffee ID.	-0.003	0.002	0.003	<b>0.032</b>	<b>0.960</b>	<b>0.168</b>	-0.004	-0.010	<b>0.044</b>	<b>0.035</b>	-0.005	0.000	<b>0.042</b>	0.016	-0.007	0.005
6 Coffee IR.	-0.004	0.000	0.004	0.012	<b>0.118</b>	<b>0.880</b>	<b>0.032</b>	<b>0.032</b>	<b>0.051</b>	0.007	-0.008	-0.003	<b>0.053</b>	<b>0.053</b>	<b>0.039</b>	<b>0.036</b>
7 Cold	<b>0.122</b>	-0.001	0.017	-0.005	-0.004	0.016	<b>0.880</b>	-0.004	<b>0.032</b>	<b>0.049</b>	-0.006	0.002	<b>0.055</b>	0.015	0.007	0.020
8 Cracker	-0.006	<b>0.073</b>	<b>0.049</b>	<b>0.035</b>	-0.008	0.021	-0.003	<b>0.818</b>	0.018	<b>0.075</b>	0.002	<b>0.061</b>	<b>0.080</b>	<b>0.112</b>	<b>0.066</b>	<b>0.092</b>
9 Laundry	<b>0.035</b>	<b>0.111</b>	<b>0.023</b>	<b>0.060</b>	<b>0.032</b>	<b>0.048</b>	<b>0.043</b>	<b>0.024</b>	<b>0.855</b>	<b>0.098</b>	0.019	<b>0.120</b>	<b>0.211</b>	<b>0.248</b>	<b>0.095</b>	<b>0.080</b>
10 Marg. Soft	0.012	<b>0.046</b>	<b>0.052</b>	<b>0.089</b>	<b>0.048</b>	0.004	<b>0.027</b>	<b>0.086</b>	<b>0.129</b>	<b>0.872</b>	0.019	<b>0.111</b>	<b>0.200</b>	<b>0.142</b>	<b>0.074</b>	<b>0.184</b>
11 Marg.Sqz.	<b>0.051</b>	-0.007	-0.004	-0.006	-0.005	-0.008	-0.006	0.001	0.009	0.017	<b>0.955</b>	0.006	<b>0.043</b>	0.011	0.013	<b>0.025</b>
12 Marg. Stick	<b>0.028</b>	<b>-0.032</b>	<b>0.064</b>	<b>0.043</b>	0.008	0.010	0.004	<b>0.075</b>	<b>0.070</b>	<b>0.060</b>	0.003	<b>0.858</b>	<b>0.104</b>	<b>0.137</b>	<b>0.138</b>	<b>0.060</b>
13 Paper Toilet	0.010	<b>0.145</b>	<b>0.061</b>	<b>0.084</b>	<b>0.048</b>	<b>0.096</b>	<b>0.042</b>	<b>0.116</b>	<b>0.245</b>	<b>0.214</b>	<b>0.032</b>	<b>0.165</b>	<b>0.787</b>	<b>0.340</b>	<b>0.153</b>	<b>0.160</b>
14 Paper Towel	0.006	<b>0.218</b>	<b>0.080</b>	<b>0.080</b>	0.018	<b>0.094</b>	0.002	<b>0.123</b>	<b>0.240</b>	<b>0.090</b>	0.007	<b>0.160</b>	<b>0.264</b>	<b>0.750</b>	<b>0.095</b>	<b>0.118</b>
15 Soap	<b>0.025</b>	<b>0.066</b>	<b>0.026</b>	<b>0.048</b>	0.011	<b>0.048</b>	0.006	<b>0.083</b>	<b>0.097</b>	<b>0.086</b>	<b>0.023</b>	<b>0.141</b>	<b>0.173</b>	<b>0.148</b>	<b>0.838</b>	<b>0.113</b>
16 Spag. Sauce	0.000	<b>0.081</b>	<b>0.096</b>	<b>0.065</b>	0.003	<b>0.036</b>	0.006	<b>0.122</b>	<b>0.079</b>	<b>0.175</b>	0.019	<b>0.091</b>	<b>0.196</b>	<b>0.131</b>	<b>0.096</b>	<b>0.863</b>

Bold categories are significantly correlated with 0.01. Bold and italic categories are significantly correlated with 0.05.

*Essays On Customization Applications in Marketing*

Table 4.16: Descriptive statistics for purchase and non-purchase weeks

A. PURCHASE WEEKS

Product	N	Price	Price per		# of Weeks	
			Volume	Promotion	on Sale	% of Prom
Allergy medicine	3557	6.649	0.271	0.240	852	0.240
Butter	3557	2.129	0.135	0.798	2840	0.798
Coffee Ground Decaf	3557	6.240	0.399	0.540	1920	0.540
Coffee Ground Regular	3557	5.678	0.275	0.755	2684	0.755
Coffee Instant Decaf	3557	5.747	0.864	0.418	1486	0.418
Coffee Instant Regular	3557	4.942	0.873	0.482	1715	0.482
Cold medicine	3557	6.901	0.274	0.464	1651	0.464
Crackers	3557	2.103	0.131	0.801	2849	0.801
Laundry	3557	6.366	0.072	0.967	3438	0.967
Margarine Soft	3557	1.525	0.155	0.941	3347	0.941
Margarine Squeeze	3557	1.567	0.134	0.681	2424	0.681
Margarine Stick	3557	1.355	0.083	0.851	3026	0.851
Paper Toilet	3557	2.144	0.550	0.985	3505	0.985
Paper Towel	3557	1.864	1.050	1.000	3557	1.000
Soap	3557	2.775	0.539	0.862	3067	0.862
Spaghetti Sauce	3557	2.351	0.087	1.000	3557	1.000

B. NON-PURCHASE WEEKS

Product	N	Price	Price per		# of Weeks	
			Volume	Promotion	on Sale	% of Prom
Allergy medicine	4689	6.639	0.271	0.244	1143	0.244
Butter	4689	2.131	0.137	0.813	3810	0.813
Coffee Ground Decaf	4689	6.292	0.401	0.555	2602	0.555
Coffee Ground Regular	4689	5.730	0.277	0.761	3567	0.761
Coffee Instant Decaf	4689	5.773	0.868	0.421	1972	0.421
Coffee Instant Regular	4689	4.972	0.880	0.485	2275	0.485
Cold medicine	4689	6.921	0.275	0.414	1940	0.414
Crackers	4689	2.119	0.132	0.754	3535	0.754
Laundry	4689	6.463	0.074	0.969	4542	0.969
Margarine Soft	4689	1.522	0.169	0.931	4367	0.931
Margarine Squeeze	4689	1.569	0.134	0.646	3029	0.646
Margarine Stick	4689	1.360	0.085	0.830	3890	0.830
Paper Toilet	4689	2.150	0.472	0.983	4608	0.983
Paper Towel	4689	1.878	1.050	1.000	4689	1.000
Soap	4689	2.785	0.587	0.849	3982	0.849
Spaghetti Sauce	4689	2.377	0.090	1.000	4689	1.000

#### **4.7.1 Model Specification and Variable Definition**

Since our model considers category incidence, we need to construct category level price and promotion variables. Category price is computed as the share-weighted average price of brands. We have only online price-promotion information. We have information on two kinds of price cut promotions: a) the product is on a “special promotion” available to all customers, b) the product is available at a special price to preferred customers who have a store card. As a promotion variable, we consider a dummy variable representing whether or not the category is on promotion. We will use category-specific intercepts to capture category preference. We denote the explanatory variables for category incidence as follows:

$$X_1 = X_1 \{ \text{Category specific intercepts, Category prices, Category promotions} \}$$

On each purchase occasion, we have price and promotion information on the product category that was chosen. If the purchased SKU is on promotion, then the category promotion variable takes the value 1 in the incidence part of the model. We also include an intercept term. The promotion variable in the expenditure part is constructed to be equal to 1 if the purchase on promotion, 0 if not. The category price variable is obtained by calculating the weighted average price per volume of SKUs in the category and weights are market share for each category. The explanatory variables for expenditure are:

$$X_2 = X_2 \{ \text{Intercepts, Category prices, Purchase is on promotion or not} \}$$

For the hierarchical regression on the price and promotion coefficients, we have only an intercept term:

$$Z_h = Z_h \{ \text{Intercept} \}$$

We have 16 categories. Our Y matrix consists of two stacked matrices:  $Y_2$  represents the observed choices of 16 categories, and  $Y_1$  is the spending for these categories across the shopping trips.

## **4.8 Results and Discussion**

We report the estimated cross-category correlation matrix for expenditure and purchase incidence across 16 categories in Table 4.17, Table 4.18 and Table 4.19. Highly correlated category pairs for expenditure are allergy and paper toilet tissue (0.772), allergy and coffee instant decaf (0.619), allergy and paper towel tissue (0.594), allergy and soft margarine (0.588), allergy and soap (0.588), allergy and stick margarine (0.568), allergy and spaghetti sauce (0.506), allergy and laundry (0.500), soft margarine and coffee instant decaf (0.523), stick margarine and soft margarine (0.501), paper toilet tissue and squeeze margarine (0.511), and finally paper towel and toilet paper (0.503). Among the margarine categories (stick, soft and squeeze), stick and soft margarine have the highest and significant correlation (0.501) for expenditure. Among the coffee categories, though the value of correlation is not very high (0.274), only the coffee instant regular and coffee instant decaf categories are significantly correlated. Expenditures of allergy and cold tablets are not significantly correlated with

### *Promotion Customization across Multiple Categories*

each other. Although expenditures of allergy tablets seem to be highly correlated with many categories' expenditures, cold tablets do not show that pattern, since this category is not correlated with any of others. We can explain this situation by considering that allergy tablets are purchased regularly, whereas customers purchase cold tablets only when necessary. Allergy tablets may thus be a store traffic driver. The paper tissue categories (towel and toilet) are significantly correlated (0.503).

Estimated purchase incidence correlations are illustrated in Table 4.18. In the paper tissue categories, purchase incidence of paper towel tissue and paper toilet tissue has one of the highest correlations, 0.615. Coffee ground decaf and coffee ground regular have the highest correlation (0.663), which is also higher compared to the correlations among the other coffee categories. The other highly correlated categories are coffee ground regular and allergy (0.515), butter and stick margarine (0.521), butter and toilet paper (0.523), butter and paper towel (0.530), toilet paper and laundry (0.515), paper towel and laundry (0.510), toilet paper and soft margarine (0.526), toilet paper and stick margarine (0.539). Chib et al. (2002) also found high correlation between toilet tissue and laundry detergents. Among the margarine categories, stick margarine and soft margarine have the highest correlation (0.559). While the expenditure of allergy and cold tablets are not correlated, purchase incidences are significantly correlated. Coincidence of coffee ground regular and coffee ground decaf, paper toilet tissue and paper towel tissue, and the butter and margarine categories can be explained from the fact that probably these category-pairs are shown on the same web page in the online shopping environment (this effect similar to the shelf effect in brick and mortar shopping). The "shelf effect" is also

observed by Chib et al. (2002) in their purchase incidence model for twelve categories. Online retailers may improve profitability by displaying high-margin products and frequently purchased products on the same web page or by presenting pop-up ads. Moreover, some high correlations between four categories in coffee and three categories in the margarine group (butter can also be included in this group) may cause consumers to purchase these categories for variety's sake rather than viewing these categories as substitutes. Although there are some negative correlations for expenditures, the purchase incidence correlations are all positive, similar to the results obtained by Chib et al. (2002).

The correlation between expenditure and purchase incidence are shown in Table 4.19. Categories that do not show a significant correlation between expenditure and purchase incidence are coffee ground decaf (0.516), coffee ground regular (0.189), cold medication (0.192), and squeeze margarine (0.369). Normally we would expect higher correlation values between purchase incidence and expenditure within a category. Expenditures on allergy medications are related to many categories' purchase incidences. Expenditure of allergy is highly related with purchase incidence of paper toilet tissue (0.812), paper towel tissue (0.649), coffee ground regular (0.637), soft margarine (0.620), spaghetti sauce (0.561), laundry detergent (0.559), coffee ground decaf (0.553) and butter (0.507). Among the coffee categories, coffee instant decaf expenditure is related with purchase incidence of coffee ground regular (0.596), coffee instant regular (0.448) and coffee ground decaf (0.459). Margarine expenditures

### *Promotion Customization across Multiple Categories*

are related to purchase incidence of butter (which is correlated with soft margarine 0.448, stick margarine 0.500 and squeeze margarine 0.416). Among margarine categories, stick margarine and soft margarine affect each other's expenditures and purchase incidence more than squeeze margarine. In the paper tissue categories, paper towel tissue and toilet tissue significantly affect each other's expenditures and purchase incidences. Expenditure of allergy tablets is significantly correlated with purchase incidence of cold tablets (0.338), however, interestingly, not the other way around (0.007). From this table, we can conclude that expenditures of coffee ground decaf, coffee ground regular, and cold tablets are mostly not significantly correlated with purchase incidence of other categories. Expenditures of cold tablets are negatively affected by the purchase incidence of the most of the categories (though these correlations are very low and therefore not significant). Expenditures of cold tablets have very low correlation values for almost all purchase incidences, which means that consumers purchase cold medicines without considering other category purchases. However, expenditures of allergy are highly related to almost all other category purchase decisions, indicating that this may be a traffic driver.

We summarize the estimated covariate effects (i.e. marketing mix, price and promotion) for the sixteen categories in Table 4.20. Modeling purchase incidence and expenditures together provides useful insights into the differing nature of price and promotion sensitivities in the choice and expenditure stages of the purchase decision. All price coefficients in the incidence part of the model are negative as expected, except for insignificant laundry price coefficients. Crackers (-0.111), squeeze

margarine (-0.156), cold (-0.130), allergy (-0.129) and spaghetti sauce (-0.147) have the highest price sensitivity of purchase incidence; allergy (-0.689), coffee instant decaf (-0.622), and cold (-0.482) have the highest price sensitivity of expenditures. We obtain a few significant positive price coefficients, which is unexpected, in the expenditure part: butter (0.078), coffee ground regular (0.267), laundry detergent (0.869). However, most of the price coefficients for both purchase incidence and expenditure are negative as expected.

According to the results in Table 4.20, promotion is not a very significant factor in the decision process of purchases of many categories in online grocery shopping except for the spaghetti sauce and toilet paper categories. The promotion coefficient of purchase incidence of spaghetti sauce is significant, equal to 1.096. In the decision process of spaghetti sauce, promotion is an important factor for purchase. Sales promotions are also an important factor in the purchase decision for paper toilet tissue (0.147). The highest promotion effects on expenditures are 18.344 of soap, 19.754 of allergy, 9.937 of squeeze margarine, 6.451 of crackers, and finally 7.592 of coffee instant regular. Spending for all margarine categories is affected by sales promotions. If consumers choose to purchase that category, sales promotion is important in influencing the decision of how much to spend (consumers may buy more than they used to, and stockpile). We can state that people's decisions regarding how much to spend are more strongly affected by sales promotions than is the decision of what category to buy. We see that promotion effects on expenditures

may be negative as well, i.e. if the price cut is large, people spend less on the product because it is cheaper, if they buy the same amount on average.

## **4.9 Optimization Results**

We report the optimization results in this section. We present the estimated objective function for each customer in Table 4.22. The objective function ( $E[\pi_i]$ ) illustrates the difference between the expected expenditure of selected categories depending on whether it is promoted or not. We estimate the expected expenditure of each category for each consumer with and without promotion, so, as shown in equation 4.11. Selecting category allocations from many different possible allocations is a combinatorial optimization problem. A similar combinatorial optimization approach to optimize the design and content of electronic communications were first applied by Ansari and Mela (2003). We generate many designs with the modified Federov design generating algorithm, which we also used to select allocations of blocks of questions to splits in the split questionnaire design problem. The promotion design with maximum promotional lift in spending for each category is accepted as the optimal promotion design. Though some revenues ( $E[\pi_i]$ ) for some customers are negative in Table 4.22, most of them are positive, as expected (expected expenditures can be negative, since if there is a high price cut, spending will decrease if the customer purchases the regular amount, as if it were not on promotion). That is, promoting a category increases consumers spending on average, but for some customers the predicted effect on spending is negative. In these cases, it may be optimal to promote fewer than the five categories

that are fixed in promotional plans. We note that in our optimal promotion design, negative revenues generally belong to customers with small shopping baskets (i.e. they spend more) or very large baskets (i.e. they spend much). In other words, managers do not need to offer promotions for customers who are already ready to spend a large amount, or for customers who spend little for their necessities during the online shopping process. In such a situation, determining the optimal number of categories to promote (i.e. number of coupons), considering the most profitable categories from many, is another important topic which must be considered.

We report the pairwise frequencies of selected categories in the final promotion design for all consumers in Table 4.23. According to this table, the promotion design offers sales promotions the most frequently for squeeze margarine (72 customers out of 133), coffee ground regular (68), cold (59), cracker (54), and finally butter (51). The least offered categories are allergy tablets (13), coffee instant decaf (17) and soap (26). Some categories are offered together more often than the others. For example, squeeze margarine and coffee ground regular are offered together 34 times, cold and squeeze margarine 32 times, cold and coffee ground regular 30 times, and toilet paper and squeeze margarine are offered together to customers (i.e. appeared together in a customized promotion plan) 26 times. On the other hand, allergy medication and coffee instant regular, stick margarine, toilet paper, and soap and coffee instant decaf, and stick margarine are offered together only once. As we notice, toilet

### *Promotion Customization across Multiple Categories*

paper and paper towel tissue are the most frequently promoted categories in the data, however our optimal customized promotion design does not suggest them to be promoted the most frequently; they appear only 13 times together.

Table 4.17: Estimated bivariate correlations of expenditure across categories

Category	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Allergy	1	0.100	0.199	0.184	0.084	0.118	0.205	0.094	0.057	0.058	0.163	0.070	0.063	0.064	0.056	0.073
2 Butter	<b>0.463</b>	1	0.138	0.148	0.089	0.134	0.189	0.073	0.044	0.041	0.131	0.053	0.032	0.037	0.051	0.040
3 Coffee GD.	0.300	0.188	1	0.166	0.179	0.144	0.168	0.161	0.143	0.143	0.155	0.164	0.136	0.164	0.172	0.155
4 Coffee GR.	0.247	0.089	0.052	1	0.180	0.150	0.132	0.159	0.120	0.146	0.117	0.147	0.112	0.161	0.148	0.155
5 Coffee ID.	<b>0.619</b>	<b>0.380</b>	0.268	0.235	1	0.137	0.158	0.124	0.133	0.091	0.149	0.134	0.091	0.127	0.112	0.108
6 Coffee IR.	<b>0.353</b>	<b>0.376</b>	0.142	0.055	<b>0.274</b>	1	0.169	0.161	0.115	0.114	0.134	0.180	0.105	0.087	0.131	0.136
7 Cold	-0.095	0.075	-0.021	0.005	0.049	-0.059	1	0.188	0.196	0.170	0.165	0.190	0.193	0.200	0.163	0.219
8 Cracker	<b>0.298</b>	<b>0.188</b>	0.011	-0.105	0.105	0.129	-0.087	1	0.088	0.084	0.135	0.085	0.072	0.075	0.076	0.067
9 Laundry	<b>0.500</b>	<b>0.325</b>	0.254	0.122	<b>0.458</b>	<b>0.295</b>	-0.179	0.138	1	0.043	0.137	0.047	0.027	0.027	0.042	0.036
10 Marg. Soft	<b>0.588</b>	<b>0.415</b>	0.144	0.122	<b>0.523</b>	<b>0.256</b>	0.135	<b>0.339</b>	<b>0.367</b>	1	0.169	0.041	0.029	0.031	0.048	0.036
11 Marg. Sqz.	<b>0.369</b>	<b>0.402</b>	0.164	0.109	0.260	<b>0.305</b>	-0.029	0.085	<b>0.409</b>	0.161	1	0.140	0.113	0.130	0.167	0.124
12 Marg. Stick	<b>0.568</b>	<b>0.478</b>	0.173	0.151	<b>0.388</b>	0.195	-0.055	<b>0.362</b>	<b>0.401</b>	<b>0.501</b>	0.208	1	0.032	0.031	0.046	0.045
13 Paper Toilet	<b>0.772</b>	<b>0.434</b>	<b>0.285</b>	<b>0.237</b>	<b>0.477</b>	<b>0.319</b>	-0.081	<b>0.261</b>	<b>0.444</b>	<b>0.451</b>	<b>0.511</b>	<b>0.463</b>	1	0.020	0.039	0.027
14 Paper Towel	<b>0.594</b>	<b>0.467</b>	0.222	0.112	<b>0.492</b>	<b>0.393</b>	-0.189	<b>0.237</b>	<b>0.408</b>	<b>0.423</b>	<b>0.374</b>	<b>0.431</b>	<b>0.503</b>	1	0.031	0.030
15 Soap	<b>0.588</b>	<b>0.409</b>	0.137	0.146	<b>0.296</b>	<b>0.349</b>	<b>-0.364</b>	<b>0.236</b>	<b>0.320</b>	<b>0.368</b>	0.160	<b>0.450</b>	<b>0.445</b>	<b>0.420</b>	1	0.044
16 Spag. Sauce	<b>0.506</b>	<b>0.365</b>	0.081	0.063	<b>0.381</b>	<b>0.302</b>	0.096	<b>0.320</b>	<b>0.327</b>	<b>0.383</b>	<b>0.305</b>	<b>0.375</b>	<b>0.419</b>	<b>0.367</b>	<b>0.365</b>	1

Lower triangle values are correlations and upper triangle (with italics) values are standard errors. Bold categories are significantly correlated with significance level 0.05.

Promotion Customization across Multiple Categories

Table 4.18: Estimated bivariate correlations of purchase incidence across categories

Category	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Allergy	1	0.145	0.136	0.113	0.136	0.131	0.129	0.145	0.094	0.112	0.122	0.114	0.102	0.103	0.111	0.120
2 Butter		1	0.063	0.056	0.098	0.088	0.108	0.048	0.031	0.039	0.107	0.047	0.022	0.024	0.040	0.031
3 Coffee GD.			1	0.072	0.108	0.104	0.119	0.069	0.059	0.071	0.122	0.065	0.050	0.054	0.060	0.055
4 Coffee GR.				1	0.108	0.097	0.131	0.080	0.052	0.051	0.119	0.055	0.039	0.036	0.055	0.051
5 Coffee ID.					1	0.110	0.166	0.132	0.089	0.084	0.162	0.122	0.080	0.079	0.109	0.117
6 Coffee IR.						1	0.131	0.119	0.059	0.087	0.142	0.109	0.054	0.053	0.090	0.078
7 Cold							1	0.144	0.081	0.090	0.153	0.123	0.072	0.076	0.090	0.106
8 Cracker								1	0.059	0.045	0.138	0.055	0.036	0.038	0.055	0.040
9 Laundry									1	0.031	0.076	0.036	0.020	0.022	0.037	0.028
10 Marg. Soft										1	0.088	0.033	0.024	0.026	0.039	0.029
11 Marg. Sqz.											1	0.093	0.075	0.067	0.116	0.091
12 Marg. Stick												1	0.024	0.028	0.040	0.036
13 Paper Toilet													1	0.015	0.027	0.021
14 Paper Towel														1	0.027	0.024
15 Soap															1	0.034
16 Spag. Sauce																1

Lower triangle values are correlations and upper triangle (with italics) values are standard errors. Bold categories are significantly correlated with significance level 0.05.

Table 4.19: Estimated bivariate correlations of purchase incidence-expenditure across categories.

Expenditure	Purchase Incidence															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Allergy	<b>0.528</b>	<b>0.507</b>	<b>0.553</b>	<b>0.637</b>	<b>0.348</b>	<b>0.438</b>	<b>0.338</b>	<b>0.436</b>	<b>0.559</b>	<b>0.620</b>	<b>0.337</b>	<b>0.595</b>	<b>0.812</b>	<b>0.649</b>	<b>0.600</b>	<b>0.561</b>
2 Butter	0.081	<b>0.966</b>	<b>0.355</b>	<b>0.360</b>	<b>0.338</b>	0.093	<b>0.332</b>	<b>0.348</b>	<b>0.443</b>	<b>0.308</b>	<b>0.497</b>	<b>0.485</b>	<b>0.510</b>	<b>0.397</b>	<b>0.406</b>	
3 Coffee GD.	0.182	0.211	0.516	0.315	0.158	0.241	0.123	0.078	0.275	0.159	0.113	0.172	<b>0.300</b>	0.243	0.136	0.107
4 Coffee GR.	0.184	0.101	0.021	0.189	0.211	0.123	0.257	-0.092	0.152	0.127	0.094	0.165	0.245	0.128	0.151	0.074
5 Coffee ID.	<b>0.412</b>	<b>0.415</b>	<b>0.459</b>	<b>0.596</b>	<b>0.672</b>	<b>0.448</b>	<b>0.347</b>	<b>0.208</b>	<b>0.519</b>	<b>0.549</b>	<b>0.248</b>	<b>0.406</b>	<b>0.519</b>	<b>0.528</b>	<b>0.302</b>	<b>0.426</b>
6 Coffee IR.	0.134	<b>0.392</b>	0.187	0.194	0.171	<b>0.491</b>	0.172	0.226	<b>0.320</b>	<b>0.273</b>	0.332	0.212	<b>0.346</b>	<b>0.420</b>	<b>0.351</b>	<b>0.328</b>
7 Cold	0.007	0.077	-0.025	-0.031	0.066	-0.043	0.192	-0.142	-0.193	0.132	-0.056	-0.054	-0.080	-0.188	-0.377	0.086
8 Cracker	0.095	<b>0.204</b>	<b>0.198</b>	<b>0.258</b>	0.001	0.162	0.048	<b>0.833</b>	0.145	<b>0.350</b>	0.161	<b>0.374</b>	<b>0.279</b>	<b>0.256</b>	<b>0.242</b>	<b>0.346</b>
9 Laundry	<b>0.326</b>	<b>0.349</b>	<b>0.279</b>	<b>0.308</b>	<b>0.193</b>	<b>0.262</b>	<b>0.191</b>	<b>0.233</b>	<b>0.960</b>	<b>0.399</b>	<b>0.464</b>	<b>0.421</b>	<b>0.467</b>	<b>0.445</b>	<b>0.345</b>	<b>0.377</b>
10 Marg. Soft	<b>0.288</b>	<b>0.448</b>	<b>0.297</b>	<b>0.419</b>	<b>0.337</b>	<b>0.268</b>	<b>0.237</b>	<b>0.404</b>	<b>0.416</b>	<b>0.970</b>	<b>0.351</b>	<b>0.522</b>	<b>0.500</b>	<b>0.467</b>	<b>0.375</b>	<b>0.435</b>
11 Marg. Sqz.	0.051	<b>0.416</b>	0.189	0.124	0.093	0.221	0.149	0.177	<b>0.425</b>	0.176	0.369	0.225	<b>0.521</b>	<b>0.403</b>	0.153	<b>0.331</b>
12 Marg. Stick	<b>0.284</b>	<b>0.500</b>	<b>0.436</b>	<b>0.409</b>	0.232	<b>0.239</b>	0.170	<b>0.448</b>	<b>0.438</b>	<b>0.535</b>	<b>0.255</b>	<b>0.946</b>	<b>0.510</b>	<b>0.471</b>	<b>0.460</b>	<b>0.417</b>
13 Paper Toilet	<b>0.235</b>	<b>0.470</b>	<b>0.421</b>	<b>0.372</b>	<b>0.233</b>	<b>0.373</b>	<b>0.240</b>	<b>0.398</b>	<b>0.486</b>	<b>0.474</b>	<b>0.319</b>	<b>0.489</b>	<b>0.984</b>	<b>0.561</b>	<b>0.444</b>	<b>0.467</b>
14 Paper Towel	<b>0.251</b>	<b>0.483</b>	<b>0.360</b>	<b>0.446</b>	<b>0.278</b>	<b>0.351</b>	<b>0.221</b>	<b>0.404</b>	<b>0.472</b>	<b>0.443</b>	<b>0.273</b>	<b>0.447</b>	<b>0.555</b>	<b>0.982</b>	<b>0.411</b>	<b>0.413</b>
15 Soap	<b>0.243</b>	<b>0.440</b>	<b>0.355</b>	<b>0.413</b>	0.183	<b>0.325</b>	0.070	<b>0.351</b>	<b>0.358</b>	<b>0.394</b>	<b>0.323</b>	<b>0.477</b>	<b>0.486</b>	<b>0.467</b>	<b>0.963</b>	<b>0.403</b>
16 Spag. Sauce	0.243	<b>0.395</b>	<b>0.307</b>	<b>0.383</b>	0.183	<b>0.268</b>	<b>0.307</b>	<b>0.388</b>	<b>0.342</b>	<b>0.412</b>	<b>0.295</b>	<b>0.421</b>	<b>0.453</b>	<b>0.408</b>	<b>0.387</b>	<b>0.973</b>

Bold categories are significantly correlated with significance level 0.05.

Promotion Customization across Multiple Categories

Table 4.20: Posterior  $\Theta$  (mean across all customers) with full  $\Lambda$

Categories	Expenditure			Incidence		
	intercept	price	promotion	intercept	price	promotion
Allergy medicine	-3.189	<b>-0.689</b>	<b>19.754</b>	-0.261	<b>-0.129</b>	0.166
Butter	-2.024	<b>0.078</b>	0.269	-1.409	<b>-0.034</b>	-0.021
Coffee Gr. Decaf	-9.983	<b>-0.156</b>	-0.923	-0.403	<b>-0.077</b>	0.103
Coffee Gr. Regular	-3.148	<b>0.267</b>	<b>-15.053</b>	0.006	<b>-0.097</b>	0.001
Coffee Ins. Decaf	4.452	<b>-0.622</b>	<b>-26.226</b>	0.632	<b>-0.054</b>	-0.018
Coffee Ins. Regular	-21.588	<b>-0.137</b>	<b>7.592</b>	-0.500	<b>-0.039</b>	0.012
Cold medicine	-9.441	<b>-0.482</b>	0.718	0.159	<b>-0.130</b>	-0.031
Crackers	-1.477	<b>-0.203</b>	<b>6.541</b>	-1.274	<b>-0.111</b>	-0.031
Laundry	-8.797	<b>0.869</b>	<b>-10.209</b>	-1.598	0.009	-0.029
Margarine Soft	-0.916	<b>-0.268</b>	-0.254	-0.639	<b>-0.085</b>	-0.100
Margarine Squeeze	2.020	<b>-0.144</b>	<b>9.937</b>	-1.545	<b>-0.156</b>	0.058
Margarine Stick	-1.929	<b>-0.140</b>	<b>-16.147</b>	-1.611	<b>-0.087</b>	-0.029
Paper Toilet	3.552	<b>-0.076</b>	0.038	0.177	<b>-0.026</b>	<b>0.147</b>
Paper Towel	1.873	<b>-0.072</b>	0.023	-0.138	-0.004	-0.536
Soap	1.631	<b>-0.106</b>	<b>18.344</b>	-0.301	-0.030	0.013
Spaghetti Sauce	1.216	<b>-0.327</b>	-0.006	-0.439	<b>-0.147</b>	<b>1.096</b>

Table 4.21: Posterior mean of standard errors of  $\Theta$ 's with full  $\Lambda$

Categories	Expenditure			Incidence		
	intercept	price	promotion	intercept	price	promotion
Allergy medicine	0.967	0.197	0.734	0.552	0.021	0.263
Butter	0.263	0.023	0.192	0.166	0.011	0.080
Coffee Gr. Decaf	1.652	0.056	0.927	0.722	0.020	0.212
Coffee Gr. Regular	0.993	0.090	1.294	0.332	0.013	0.111
Coffee Ins. Decaf	1.249	0.098	1.677	0.605	0.009	0.236
Coffee Ins. Regular	0.444	0.047	0.615	0.598	0.007	0.280
Cold medicine	0.589	0.041	1.111	0.844	0.029	0.149
Crackers	0.522	0.046	0.955	0.476	0.035	0.152
Laundry	0.718	0.093	1.799	0.356	0.045	0.121
Margarine Soft	0.614	0.046	0.384	0.489	0.031	0.140
Margarine Squeeze	0.889	0.045	0.999	0.421	0.036	0.217
Margarine Stick	0.519	0.044	1.333	0.310	0.036	0.129
Paper Toilet	1.597	0.035	0.055	0.592	0.012	0.073
Paper Towel	2.482	0.030	0.094	0.418	0.005	0.420
Soap	0.453	0.012	0.695	1.050	0.018	0.072
Spaghetti Sauce	0.636	0.072	0.151	0.879	0.047	0.546

Table 4.22 Optimization results

Consumer	E[p <sub>i</sub> ]	Consumer	E[p <sub>i</sub> ]	Consumer	E[p <sub>i</sub> ]	Consumer	E[p <sub>i</sub> ]
1	-2.923	35	0.878	69	-0.728	103	-0.062
2	3.349	36	-1.589	70	-2.297	104	-1.018
3	0.621	37	-1.133	71	-1.799	105	2.303
4	1.880	38	-2.813	72	1.726	106	-2.670
5	0.212	39	-0.675	73	1.707	107	-3.201
6	-0.618	40	0.354	74	-0.511	108	-0.299
7	0.764	41	0.958	75	-0.300	109	-0.597
8	-3.687	42	1.714	76	-1.970	110	-0.995
9	1.225	43	-0.737	77	-2.033	111	0.504
10	-2.926	44	0.574	78	0.480	112	1.538
11	1.444	45	-3.032	79	0.709	113	-1.576
12	-0.627	46	-1.159	80	1.296	114	-0.208
13	-1.487	47	-0.379	81	-2.495	115	-1.053
14	2.769	48	-0.340	82	1.330	116	0.590
15	-1.826	49	0.778	83	1.180	117	0.763
16	-0.960	50	0.832	84	3.072	118	0.723
17	0.032	51	-0.016	85	-2.370	119	-3.069
18	-0.586	52	-0.607	86	1.392	120	1.584
19	-0.360	53	0.976	87	-0.391	121	-3.365
20	0.581	54	0.067	88	-0.740	122	-0.327
21	0.410	55	2.538	89	0.880	123	-0.439
22	-0.652	56	-0.258	90	-1.698	124	0.308
23	1.084	57	-1.917	91	-2.312	125	0.764
24	-0.306	58	-2.967	92	2.218	126	2.249
25	-1.185	59	-0.832	93	3.136	127	-0.753
26	1.151	60	-3.235	94	0.063	128	-1.635
27	-3.463	61	0.099	95	2.176	129	1.071
28	-0.781	62	-0.490	96	-0.485	130	0.520
29	-0.444	63	-2.274	97	3.175	131	-0.288
30	-3.347	64	1.707	98	-1.401	132	-0.709
31	0.155	65	0.821	99	-1.475	133	4.001
32	2.284	66	1.108	100	-2.470		
33	1.305	67	0.212	101	-0.751		
34	1.461	68	0.486	102	-0.231		

Note: E[ $\pi_i$ ] is the difference between expenditure of a category if promoted or not promoted, calculated using price per volume

Table 4.23: Frequencies of offers across categories

<b>Category</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>	<b>11</b>	<b>12</b>	<b>13</b>	<b>14</b>	<b>15</b>	<b>16</b>
<b>1</b> Allergy	13															
<b>2</b> Butter	3	51														
<b>3</b> Coffee GD.	4	9	36													
<b>4</b> Coffee GR.	6	25	11	68												
<b>5</b> Coffee ID.	2	4	4	9	17											
<b>6</b> Coffee IR.	1	14	8	12	2	30										
<b>7</b> Cold	3	18	13	30	4	7	59									
<b>8</b> Cracker	4	15	15	25	5	10	25	54								
<b>9</b> Laundry	5	14	11	16	4	4	9	11	33							
<b>10</b> Marg. Soft	4	17	8	18	4	7	17	5	5	38						
<b>11</b> Marg.Sqz.	3	22	18	34	11	19	32	25	17	15	72					
<b>12</b> Marg. Stick	1	12	4	12	4	4	8	8	6	7	16	28				
<b>13</b> Paper Toilet	1	14	14	19	5	11	21	23	11	13	26	9	49			
<b>14</b> Paper Towel	7	18	7	18	5	7	18	18	9	12	20	10	13	46		
<b>15</b> Soap	6	8	9	14	1	6	9	10	5	10	8	1	5	4	26	
<b>16</b> Spag. Sauce	2	11	9	23	4	8	22	17	5	10	22	10	11	18	8	45

## 4.10 Conclusion

In this chapter, we focus on customization of promotions across multiple categories. We have a model which allows estimating the effects of promotions what categories consumers purchase and how much they spend for each category. We use a hierarchical Bayes multivariate type-2 tobit model for that. We used the Gibbs sampling to estimate model parameters. After estimating the model, we approach the design of customized promotion plan design with the Bayesian decision approach. Our objective function in the optimization problem is revenue, which is a function of the expected spending of a category when we do or do not promote. We estimate the objective function in each MCMC chain, which allows us to integrate out the uncertainty in the parameters. After we estimate the expected promotional lift in expenditure for each customer for

each category, we use the Federov design generating algorithm to choose the best among many possible customized promotion designs. The designs with the maximum expected promotional lift in spending for the selected categories are the optimal customized promotion designs. We found an optimal promotion design for each customer, and thus customization was successfully performed with the Bayesian decision framework, although it appeared that for some customers the chosen numbers of promotions (five) results in a decrease in revenue, so that this subsection of customers, less than five promotions may be optimal.

The estimated correlation matrix illustrates that purchase incidence and expenditure should be modeled together. Interdependence between purchase incidence and expenditure between categories should be examined more carefully. Taking these cross-correlations into account helps to create more efficient marketing strategies, including cross-selling strategies, setting optimal prices, creating more efficient shopping websites, creating better shelf designs in stores, creating more successful online or in store coupon strategies, and designing better customization strategies.

Empirical results illustrate that sales promotion is an important factor for the decision of how much to spend. Consumers may buy more than they used to and stockpile under sales-promotions. People's decisions of how much to spend is more affected by sales promotions compared to the decision of what category to buy. Promotion effects of expenditures may also be negative, i.e. if the price cut is large, people spend less on the product because it is cheaper, if sales volume does not increase. We

### *Promotion Customization across Multiple Categories*

observe effects of shelf layout in online shopping, similar to in brick and mortar stores. Since some categories are presented on the same web page, they are purchased together more frequently. For example, more frequent coincidence of coffee ground regular and coffee ground decaf, paper toilet tissue and paper towel tissue, and butter and margarine categories can be explained by this. We found sales promotions to be most important for the purchase decision of spaghetti sauce. Price effects are significant in almost all categories for both purchase incidence and expenditure decisions. The expenditures on coffee ground decaf, coffee ground regular, and cold tablets are mostly not significantly correlated with purchase incidence of other categories. Expenditures of allergy tablets are highly related to purchase decisions in almost all other categories, indicating that this may be a traffic driver.

The model used in this chapter can be potentially improved in four ways: 1. Include consumer budget constraints in the estimation and the optimal allocation of promotions. 2. Include dynamics in the price and promotion parameters (state-space approach) so that the optimal allocation of promotion varies over time and is dependent on reactions to the most recently promoted categories. 3. Build a brand-choice model on top of the category expenditure and incidence model, so that we also know what brands to promote in each category. 4. Reformulate the model into a purchase quantity and incidence model, considering budgetary constraints.





# Chapter 5

## Conclusion and Discussion

### 5.1 Introduction

The primary objective of this thesis is to develop and validate new methodologies to improve the collection of data and the effectiveness of promotion customization. For this purpose, we use the Bayesian approach in every stage of problem solving, inference, estimation and decision making. This thesis contains two essays. The first essay deals with how to improve collection of data. We recommend using split questionnaires for long questionnaires, common in marketing, and develop a methodology to design optimal split questionnaires. The second essay is a cross-category promotion customization problem. We fit and estimate a model across multiple categories, which allows cross-category promotion strategies. We customize optimal promotion design with a combinatorial optimization approach. In this chapter, the main conclusions are summarized and further research is suggested.

### 5.2 Summary and Conclusions

In the first chapter, we explained the term “customerization”. Marketing managers need new online marketing strategies such as new methods of interacting with customers because of the migration of marketing to the online environment. Customization and customerization are two different concepts that should not be confused (Wind and Rangaswamy, 2001). Customerization indicates the customer’s individual likes and dislikes which

are placed at the center of every stage of the marketing process, rather than only tailoring the offering. Customerization can be summarized as merging strategies of one-to-one marketing, personalization, targeting and mass-customization. Successful customerization strategies should combine supply and demand sides. From this perspective, the possibilities of customizing marketing mix instruments from the seller and buyer side are explained in Chapter 1. To develop new customerization strategies, we need customer information (likes, dislikes, lifestyles, purchase habits, some background variables etc.), which is critical for identifying, differentiating, and interacting with customers. This information can be collected with questionnaires. Better tools for data collection and better models for customization will help managers or market researchers make better decisions. The Bayesian framework will prove to be useful. Two applications are presented on collecting consumer data on soft variables such as lifestyles or consumer satisfaction in Chapter 3 and cross-category promotions in Chapter 4.

In the second chapter, we present why the Bayesian approach is particularly appropriate to the decision orientations of marketing problems. In the Bayesian approach, all available information is used to reduce the amount of uncertainty which is present in an inferential or decision-making problem. The main reason for the increased usage of Bayesian methods in marketing in the last decade is not only the increasing capacity of computers and the success of MCMC algorithms to solve complex marketing problems, but also the reliance on the characteristics of

### *Conclusion and Discussion*

marketing data, the necessity to approach marketing problems as a decision problem and the flexibility and robustness of Bayesian methods. Marketing models with latent variables, missing data, mixed outcome data, heterogeneity of coefficients, nonlinearity, discrete data and more, are easy to estimate with in the Bayesian framework. Since the Bayesian paradigm uses all information and merges prior information with observed data to estimate models (updating information), the Bayesian paradigm is optimal for decision problems in marketing. The Bayesian decision process considers the estimation or model uncertainty in the complete process of problem solving.

In the third chapter, we focus on split questionnaires to collect data instead of using the more typical long questionnaires (i.e. more than 20 minutes) in marketing, since they offer the potential to obtain higher quality information from respondents faster and at a substantially lower cost. In split questionnaires, different respondents respond different parts of the questionnaire. That is, we have different versions of the questionnaire that are shorter than the whole questionnaire. After generating different split questionnaires and administrating them to respondents, we impute data for the missing parts using the other people's responses to those missing parts. In the end, we obtain almost the same information with split questionnaires as with complete lengthy questionnaires, but in a shorter time with less cost and obtaining better quality responses (less item nonresponse, higher response and more accurate responses). We propose a methodology to generate split questionnaire versions based on some prior information and for this we use optimal experimental design methods. We generate many designs with the modified Federov algorithm to search

over the design space using the Kullback-Leibler distance as a design criterion, and illustrate that good designs are feasible. We present synthetic data results for algorithm performance, real data results for statistical efficiency (i.e. we show that we obtain almost the same information with split questionnaires compared to full questionnaires) and field study results that reveal behavioral efficiency. The statistical and behavioral efficiency of split questionnaire designs are shown by comparing them to full questionnaires or questionnaires constructed with ad-hoc methods.

In the fourth chapter, the cross-category promotion design problem is presented. Currently, many multicategory models are restricted to a smaller number of categories, and the main purpose in these studies is to understand any type of demand relationship across product categories (substitution, complementarity or independence). Retailers can use cross-category relatedness for delivery of point-of-purchase materials, cross-category coupons, creative store layout, and online feature ad design. In marketing, we need better models to understand consumers' multicategory preferences. In our application, we consider not only category interdependencies, but also purchase incidence and expenditure interdependencies using a hierarchical Bayes type-2 tobit model. We approach the problem of optimally promoting a limited set of categories as a combinatorial optimization problem and use a design generating algorithm to generate many promotion designs to find the optimal one. The objective function is the maximum profit change if the category is promoted for the selected categories. The approach presented here can be applied to

### *Conclusion and Discussion*

other possible marketing offerings, which can be a product, service, combination of product and service, or a bundle of products and/or services. During the model estimation and optimization, we use a Bayesian approach which allows considering estimation uncertainty and parameter uncertainty. In decision processes in marketing, the degree of uncertainty in making decisions needs to be communicated effectively to managers. As we know, managers are generally risk-averse, and analysis involving exogenous variables can generate predictions that vary greatly in their precision. From this perspective, predictive uncertainty should be communicated to managers, and thus they can favor decisions that correspond to more certain predictions, or they can collect additional information to further reduce uncertainty.

### **5.3 Limitations and Future Research**

Despite the many developments that have already taken place in conjoint questionnaire design literature, there is a need to collect better quality data with long questionnaires in surveys in marketing. Successful customization strategies depend on better models and better optimization algorithms. We propose a combinatorial optimization approach, which allows optimal assignment of promotions across multiple categories. In this section, we consider those avenues, as well as limitations and possible (or planned) extensions that are related to the research presented in Chapter 3 and 4.

### **5.3.1 Split Questionnaires**

First, we present some methodological limitations. We assume a multivariate normal distribution for variables in questionnaire in Chapter 3. The SQD method could be extended to accommodate binomial data or mixtures of categorical and continuous data, based on the general location-scale model (Olkin, 1961) to enable one to optimally split and impute questionnaires in a wider variety of questionnaire design problems. Although we mention how we can extend the proposed questionnaire design problem to the mixed data case with the general location model, we have not included results into this thesis. The extension of the proposed method in design and imputation stage for the mixed data case is in progress.

One limitation of this research could be our multiple imputations based on the multivariate normal distribution. We intend to extend the imputation procedure to the mixed data case, on which there is already existing literature. Although we imputed data without considering individual heterogeneity, there are some studies that consider multiple imputation at the individual level. Gelman et al. (1998) propose a multiple imputation procedure, which assumes an imputation model that allows one to include covariates on the individual and the survey. In this procedure, the individual heterogeneity enters into the imputation procedure through the multivariate normal data model, with a common covariance matrix and differing mean vectors concerning the survey levels at the individual level.

### *Conclusion and Discussion*

We used a greedy algorithm to design within block designs, however the sensitivity of a within block design to prior information is relatively high with this algorithm. One would prefer to generate those designs with different algorithms that reduce the sensitivity of this design to prior information. Therefore, better design algorithms for the within block designs are an avenue for future research.

One of the problems of using prior information to construct split designs is that we are still using a full questionnaire from a pilot study or sub sample. Therefore, any undesirable response styles from a full questionnaire may still affect our generation of split questionnaire designs. However, we can prevent this problem by checking the response pattern of subjects in a full questionnaire. We can eliminate or correct the questions with bad response styles from a full questionnaire, or we can develop more extensive models that take response styles into account.

It would also be of future interest to develop dynamic algorithms to optimize questionnaire designs. SQD methods are arguably important tools in web-based surveys, online panel surveys and pop-up questionnaires (Comley, 2000). Here, respondent burden is an even more important issue, and our method could be extended to allow dynamic updating of the split questionnaire design for each respondent as more data comes in. Using some past information from subjects, one can then very quickly and efficiently customize the split questionnaire to individual customers. We believe for “real time” marketing decisions, online questionnaires can be an important tool in the future. The main limitation of an online survey is its length. Online surveys should be short. Fram and Grady (1995) found

consumers unwilling to respond to lengthy surveys administered online. Principals at NFO Research, Inc. also report that participation rates drop dramatically when online surveys become long (e.g., more than 40 items). As a result, questionnaire constructs and concepts must be captured parsimoniously in interactive surveys, and split questionnaires methodology can be applied to design online surveys. In fact, online questionnaires already have become a main interest in conjoint analysis. Since these questionnaires can be used for pricing and new product development analysis. Sometimes one may need to merge survey data and behavior data. For example, De Bruyn et al. (2005) did a study on online conjoint questionnaires, in which they used survey data questions and behavioral data to collect consumer preferences using shorter conjoint questionnaires. Modifications of our method to “real-time” decisions are also possible. For instance, we can design split questionnaires in two stages. First, we may need to segment customers with certain questions (or common questions) at an initial stage, and then customize questionnaires based on these segments to collect data more efficiently in the second stage. We may have some prior information to understand which questions can be used to classify respondents to segments. When respondents start to respond the questions, this prior information can be updated at the second stage using the Kullback-Leibler distance as an information statistic at each stage. The order of the question can then be decided for each individual based on this information statistic. At any moment during the questionnaire, we can know

### *Conclusion and Discussion*

the probability of the respondent belonging to each of the segments using the finite mixture model (Kamakura and Wedel, 1998).

Another possibility for future research is to study the optimal sample size for each version of the split questionnaire. Although we distributed each distinct split (versions of the questionnaire) to respondents evenly in our application, “sample size” for each questionnaire is an important issue one should consider. Each split should be distributed to a sufficient number of respondents for validity. Then the question becomes how can we estimate efficient sample sizes for each split? Are there some versions of the questionnaire that need more subjects, while others do not? Should sample size depend on the number of questions in the questionnaire, or on the information content of each split questionnaire? We think these questions are of interest for future research.

Although we do not need common questions to design split questionnaires, using them may increase the efficiency of imputations. In split questionnaire design where certain common questions are contained in both versions of the questionnaire, attention should be given to the ordering and positioning of these common questions to reduce potential response bias and/or carryover effect. For population surveys conducted regularly over time, variables such as gender, age, background, etc. can easily be conceived as common variables when the change of population dynamics over a certain time period can be ignored. If it is necessary to ask certain questions to every respondent, we may use them as common variables.

In the future, we probably will see more applications of split questionnaires in media and purchasing behavior panel surveys. Ressler (2002) claims that questionnaires used in the television measurement panel and the purchasing behavior panel can be reorganized to create suitable blocks of variables. Currently used methods for this purpose, such as data fusion, depend on the conditional independence assumption and therefore suffer from identification problems. On the other hand, by overlapping blocks of questions, split questionnaires can provide a solution to the identification problem by overcoming the conditional independence assumption.

There are several issues to be resolved in future research. We generated SQDs using the number of splits as an external constraint. SQDs can be easily generated under different constraints that arise in practical applications, but the performance of the SQD under such constraints remains to be investigated. As an illustration, we investigated the effect of the constraint of five blocks for each split in our empirical application. We obtain a much larger reduction of the number of questions with the five-block constraint, however, at the same time, the percentage of missing information increases and the performance of the SQD decreases in an absolute sense (although its performance relative to the RQD seems to improve). Our purpose is to eliminate questions at a minimum cost of information loss. Therefore, considering all possible splits without any constraint may be more desirable, since there is more opportunity to borrow information between blocks, which increases the efficiency of the

### *Conclusion and Discussion*

imputations. Nevertheless, in applications, additional constraints, such as the number of blocks of questions (and indirectly the total number of questions to ask in a survey) in each split of the questionnaire may be important. These kinds of constraints can ensure less respondent fatigue, which is very useful for practitioners. In short, the pay-off of reducing respondent burden versus information loss is another important topic for future research.

#### **5.3.2 Customization**

First, we discuss a few methodological issues. The main limitation in the current approach is the design generating algorithm used, since it is limited to twenty categories. As the number of categories becomes large, this approach will become infeasible. We can possibly increase this number using different design generating algorithms to find optimal product promotions. Additionally, in case we have many explanatory variables, we may consider estimating the error covariance of individual level marketing mix parameter estimates with a factor approach. We can model the large covariance matrix using a parsimonious factor analytic model.

There are a number of limitations of the model used in Chapter 4, all of which provide new directions for future research. The model developed ignores brand choice. We may need a better model to accommodate households' brand choice decisions within each product category. For that, we can use the multinomial logit (or probit) model for households' conditional brand choice within each product category, in combination with the multivariate probit model for category purchase incidence. After including brand choice in the category expenditure and incidence model,

we should also consider it in the optimization problem. Then our design problem can be defined as choosing a certain number of categories to promote from among many, and from these categories, which brand should be selected for a sales promotion. There are two possible approaches for the brand promotion allocation problem. The first one is to use a two-stage design generating algorithm. At the first stage, we can decide which categories to promote, and at the second stage, we can choose a brand to promote from the selected categories. The second approach is that we can select a brand from each category using the greedy algorithm that we also used to design within block designs.

Our model can be improved by including the consumer's budget. We expect a high degree of correlation between purchase incidences and the consumer's shopping budget. In fact, budget constraints may cause cross-category dependencies as well. However, the budget constraint is unobserved. Assuming a utility function for each category, we may assume that a customer chooses to allocate his/her dollars across categories by selecting that allocation that maximizes this utility function. Budget constraints in the context of searching for cross-category effects were previously applied by Song and Chintagunta (2003).

Another direction would be to include dynamics (i.e. state dependence) into price and promotion effects across categories. Then, the allocation of promotion varies over time and depends on the reactions to the categories promoted last. State dependence can enter the model with lagged purchase incidence variables, or with time varying marketing mix variables.

### *Conclusion and Discussion*

There are some studies within the context of random utility models that investigate the effects of a household's current choice on its future choices. Seetharaman, Ainslie and Chintagunta (1999) investigate similarities and differences in household state dependence behavior across multiple categories, relying on the effects of household and category variables. They study cross-category state dependence effects across five categories (ketchup, peanut butter, margarine, toilet tissue, and tuna). They used a multinomial probit model for choice within a category and a Bayesian variance components model (Ainslie and Rossi, 1998) for the covariation of household response parameters across categories. Optimal promotional allocation based on that approach would be dynamic and change over with time.

The composition of the shopping basket is one of the many multicategory choice phenomena that are encountered by consumers. We can use our approach to model consideration formation sets of consumers. Sometimes, due to limited information processing capacity, consumers narrow their options to simplify their decision task. We use the term "consideration set" in marketing, which covers the brands consumers consider acceptable for the next purchase. Explicitly, consideration set refers to the set of brands (a subset of all the brands in the product category) between which a consumer makes an explicit utility comparison or cost benefit trade-off before she makes her brand choice decision. We can extend our approach to include consideration sets. From online transaction data, we observe that in general consumers purchase only a few brands in the category, and sometimes never purchase some of the categories. Each household choice set can be modeled by taking the set of

all possible subsets of the available brands which are purchased by the consumer, and assigning a household specific probability distribution on each set (DeSarbo et al., 1995, Chiang et al., 1999). Using this approach, we can generate the set of all possible subsets of brands, using the design generating algorithm for each consumer, based on observed a priori category and brand choices. Such a model would considerably simplify the optimal promotional design problem. Since for each customer, we don't need to use all categories and brands in the category, but only the ones actually considered, our optimization problem can be simplified, and promotion designs can be possible across a large number of categories.

In this thesis, we studied two different topics which can be useful for solving internet marketing problems. Nowadays, the new way of communication to customers is the Internet, and for this reason the importance of developing methodologies in an online environment is increasing. We believe that questionnaire design methods, especially split questionnaire design, will receive more interest in marketing research in the future in order to collect data more efficiently (i.e. cheaper and faster) and with better quality. Moreover, this topic presents many avenues for future research. Customization is an important topic for marketing academics, market researchers and especially marketing managers. In addition to its substantive applications in marketing, there will be increasing opportunities for research on the topic.



## References

- Adams, La Mar L. and Darwin Gale (1982), "Solving the Quandary Between Questionnaire Length and Response Rate in Educational Research," *Research in Higher Education*, 17 (3), 231-240.
- Adiguzel, Feray and Michel Wedel (2004), "The Design of Split Questionnaires," Working Paper, University of Michigan
- Ainslie, Andrew and Peter E. Rossi (1998), "Similarities in Choice Behavior across Product Categories," *Marketing Science*, 17(2), 91-106
- Allenby, Greg M. and Peter Lenk (1995), "Reassessing Brand Loyalty, Price Sensitivity, and Merchandising Effects on Consumer Brand Choice," *Journal of Business Economic Statistics*, 13, 281-289
- Allenby, Greg M., Neeraj Arora, and James L. Ginter (1995), "Incorporating Prior Knowledge into the Analysis of Conjoint Studies," *Journal of Marketing Research*, 32, 152-162
- Allenby, Greg M. and James L. Ginter (1995), "Using Extremes to Design Products and Segment Markets," *Journal of Marketing Research*, 32, 392-403
- Allenby, Greg M. (1998), "On the Heterogeneity of Demand," *Journal of Marketing Research*, 35, 384-389
- Andrews, Rick, Asim Ansari A., and Imran S. Currim (2002), "Hierarchical Bayes versus Finite Mixture Conjoint Analysis Models: a Comparison of Fit, Prediction, and Partworth Recovery," *Journal of Marketing Research*, 39, 87-98
- Anjos, Miguel F., Russel C.H. Cheng, and Christine S.M. Currie (2005), "Optimal Pricing Policies for Perishable Products," *European Journal of Operational Research*, 166, 246-254
- Ansari, Asim, Skander Essegaier, and Rajeev Kohli (2000), "Internet Recommendation Systems," *Journal of Marketing Research*, 37 (August), 363-375

## References

- Ansari, Asim and Carl Mela (2003), "E-customization," *Journal of Marketing Research*, 40(3), 131-145
- Bar-Hen, Avner and J. J. Daudin (1995), "Generalization of the Mahalanobis Distance in the Mixed Case," *Journal of Multivariate Analysis*, 53, 332-342.
- Barnard, John, Robert E. McCulloch, and Xiao-Li Meng (2000), "Modeling Covariance Matrices in terms of Standard Deviations and Correlations, with Applications to Shrinkage," *Statistica Sinica*, 10(4), 1281-1311
- Bean, Andrew G. and Michael J. Roszkowski (1995), "The Long and Short of It: When Does Questionnaire Length Affect Response Rate," *Marketing Research*, 7 (1), 21-26.
- Berdie, Douglas R. (1989), "Reassessing the Value of High Response Rates to Mail Surveys," *Marketing Research*, 1 (3), 52-64.
- Berger, James O. (1985), *Statistical Decision Theory and Bayesian Analysis*, New York: Springer
- Bertsimas, Dimitris J. and Adam J. Mersereau (2003) "A Learning Approach to Customized Marketing," working paper, Stanford University
- Blattberg, Robert C. and Edward I. George (1991), "Shrinkage Estimation of Price and Promotional Elasticities: Seemingly Unrelated Equations," *Journal of American Statistical Association*, 86, 304-315
- Boatwright, Peter, Robert E. McCulloch, and Peter E. Rossi (1999), "Account-level modeling for trade promotion: An Application of a Constrained Parameter Hierarchical Model," *Journal of American Statistical Association*, 94, 1063-1073
- Bolton, Ruth N. (1989), "The Relationship between Market Characteristics and Promotional Price Elasticities," *Marketing Science*, 8(2), 153-169

## References

- Bradlow, Eric T. and David C. Schmittlein (2000), "The Little Engines that could: Modeling the Performance of World Wide Web Search Engines," *Marketing Science*, 19 (1), 43-62
- Bradlow, Eric T., Ye Hu, and Teck-Hua Ho (2004), "Learning-Based Model for Imputing Missing Levels in Partial Conjoint Profiles," *Journal of Marketing Research*, 41(4), 369-381
- Brockett, Patrick L., Perry Haaland, and Arnold Levine (1981), "Information Theoretic Analysis of Questionnaire Data," *IEEE Transactions on Information Theory*, 27(4), 438-446
- Bucklin, Randolph E., James M. Lattin, Asim Ansari, David Bell, Eloise Coupey, John D. C. Little, Carl Mela, Alan Montgomery, and Joel Steckel (2002), "Choice and the Internet: from Clickstream to Research Stream," *Marketing Letters*, 13(3), 245-258
- Bucklin, Randolph E. and Catarina Sismeiro (2003), "A Model of Web Site Browsing Behavior Estimated on Clickstream Data," *Journal of Marketing Research*, 40 (August), 249-267
- Casella, George and Roger L. Berger (1990), *Statistical Inference*, California: Brooks/Cole
- Chaloner Kathryn (1996), "The Elicitation of Prior Distributions," to appear in *Bayesian Biostatistics*, eds. D. Berry and D. Stangl, New York: Marcel Dekker.
- Chiang, Jeongwen, Siddhartha Chib, and Chakravarthi Narasimhan (1999), "Markov Chain Monte Carlo and Models of Consideration Set and Parameter Heterogeneity," *Journal of Econometrics*, 89, 223-248
- Chib, Siddhartha (1992), "Bayes Inference in the Tobit Censored Regression Model," *Journal of Econometrics*, vol. 51, issue 1-2, 79-99
- Chib, Siddhartha, Seethu P.B. Seetharaman, and Andrei Strijnev (2002), "Analysis of Multi-category Purchase Incidence Decisions Using IRI Market Basket Data," *Econometric Models in Marketing*, Vol. 16, 57-92

## References

- Chintagunta, Pradep and Sudeep Haldar (1998), "Investigating Purchase Timing Behavior in Two Related Product Categories," *Journal of Marketing Research*, 35, 45-53
- Comley, Pete (2000), "Pop-up surveys. What works, what doesn't work and what will work in the future," *Proceedings of the ESOMAR worldwide Internet conference Net Effects 3*, Publication series – Vol. 237.
- Cook, Dennis R. and Christopher J. Nachtsheim (1980), "A Comparison of Algorithms for Constructing Exact D-optimal Designs," *Technometrics*, 22(August), 315-324.
- Cyert, Michael R. and Morris H. DeGroot (1987), *Bayesian Analysis and Uncertainty in Economics*, New Jersey: Rowman and Littlefield
- De Finetti, Bruno (1970), *Theory of Probability*, New York: Wiley
- De Bruyn, Arnaud, John C. Liechty, Eelko K.R.E. Huizingh, and Gary L. Lilien (2005), "Offering Online Recommendations to Impatient, First-Time Customers with Conjoint Based Segmentation Trees," in MSI Working Paper Series (ref. 05-103).
- DeSarbo, Wayne S. and Kamal Jedidi (1995), "The Spatial Representation of Heterogeneous Consideration Sets," *Marketing Science*, 14(3), 326-243
- DeSarbo, Wayne S. and Jungwhan Choi (1999), "A Latent Structure Double Hurdle Regression Model for Exploring Heterogeneous Consumer Search Patterns," *Journal of Econometrics*, 89 (1-2), 423-455
- Dillman, Don A. (1991), "The Design and Administration of Mail Surveys," *Annual Review of Sociology*, 17, 225-249.
- Dillman, Don A., Michael D. Sinclair, and Jon R. Clark (1993), "Effects of Questionnaire Length, Respondent-Friendly Design, and a Difficult

## References

- Question on Response Rates for Occupant-Addressed Census Mail Surveys," *Public Opinion Quarterly*, 57(3), 289-304.
- Dorfman, Jeffrey H. (1997), *Bayesian Economics through Numerical Methods*, New York: Springer-Verlag
- Edwards, Yancy D., Greg M. Allenby (2003), "Multivariate Analysis of Multiple Response Data," *Journal of Marketing Research*, 40(August), 321-334
- Elrod, Terry (2005), "Bayesian Modeling of Human Behavior Using WinBUGS," a tutorial, University of Alberta
- Fader, Peter and Leonard M. Lodish (1990), "A Cross-category Analysis of Category Structure and Promotional Activity for Grocery Products," *Journal of Marketing*, 54,52-65
- Federov, Valeri V. (1972), *Theory of Optimal Experiments*, translated and edited by W.J. Studden and E.M. Klimko, New York: Academic Press
- Fox, Edward J., Alan L. Montgomery, and Leonard M. Lodish (2004), "Consumer Shopping and Spending across Retail Formats," *Journal of Business*, 77(2), 825-860
- Fram, Eugene H. and Dale B. Grady (1995), "Internet buyers: will the Surfers become buyers?," *Direct Marketing*, October, 63-65
- Gelfand, Alan E. and Adrian Smith (1990), "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, Vol. 85, 398-409.
- Gelman, Andrew, Gary King, and Chuanhai Liu (1998), "Not Asked, Not Answered: Multiple imputation for Multiple Surveys," *Journal of the American Statistical Association*, 93(443), 846-857
- Geman, Stuart and Donald Geman (1984), "Stochastic Relaxation, Gibbs Distributions, and Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6, 721-741

## References

- Giesbrecht, Francis (2004), *Planning, Construction, and Statistical Analysis of Comparative Experiments*, Wiley: New Jersey.
- Gilks Wally R., Sylvia Richardson, and David J. Spiegelhalter (1996), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, UK.
- Gilula, Zvi, Robert E. McCulloch, and Peter E. Rossi (2006) "A Direct Approach to Data Fusion," *Journal of Marketing Research*, 43(1), 73-83
- Good, Irving J. (1969), "Split Questionnaires I," *The American Statistician*, 23(4), 53-54
- Good, Irving J. (1970), "Split Questionnaires II," *The American Statistician*, 24(2), 36-37
- Gooley, Christopher G. and James Lattin (2000), "Dynamic Customization of Marketing Messages in Interactive Media," Research paper No. 1664, Graduate School of Business, Stanford University, Stanford, CA
- Gupta, Sunil (1988), "Impact of Sales Promotions on When, What, and How Much to Buy," *Journal of Marketing Research*, 25(November), 342-355
- Haaland, Perry and Patrick L. Brockett (1979), "A Characterization of Divergence with Applications to Questionnaire Information," *Information and Control*, 41, 1-8
- Heberline, Thomas A. and Robert Baumgartner (1978), "Factors Affecting Response Rates to Mailed Questionnaires: a Quantitative Analysis of the Published Literature," *American Sociological Review*, 43(4), 447-462.
- Hermkens, Piet L. J. (1983), *Oordelen over de Rechtvaardigheid van Inkomens* [In Dutch:Judgements on the Fairness of Income], Amsterdam: Kobra

## References

- Herzog, Regula A. and Jerald Bachman (1981), "Effects of Questionnaire Length on Response Quality," *Public Opinion Quarterly*, 45 (4), 489-504.
- Huber, Joel, Dick R. Wittink, John A. Fiedler, and Richard Miller (1993), "The Effectiveness of Alternate Preference Elicitation Procedures in Predicting Choice," *Journal of Marketing Research*, 30 (1), 105-114.
- Joel, Huber and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, 33 (August), 307-317
- Kalyanam, Kirthi (1996), "Pricing Decision under Demand Uncertainty: A Bayesian Mixture Model Approach," *Marketing Science*, 15, 207-221
- Kalyanam, Kirthi and Thomas S. Shively (1998), "Estimating Irregular Pricing Effects: A Stochastic Spline Regression Approach," *Journal of Marketing Research*, 35, 16-29
- Kamakura, Wagner and Michel Wedel (1997), "Statistical Data Fusion for Cross-Tabulation," *Journal of Marketing Research*, 34, 485-498
- Kamakura, Wagner and Michel Wedel (1998), *Market Segmentation: Conceptual and Methodological Foundations*, International Series in Quantitative Marketing, Kluwer: the Netherlands.
- Kamakura, Wagner, and Michel Wedel (2000) "Factor Analysis and Missing Data," *Journal of Marketing Research*, 37, 490-498
- Kim, Jaehwan, Greg M. Allenby, and Peter E. Rossi (2002), "Modeling Consumer Demand for Variety," *Marketing Science*, 21, 223-228
- Kish, Leslie (1965), *Survey Sampling*, John Wiley&Sons: New York
- Krishnamurti, Lakshman and S. J. Raj (1988), "A Model of Brand Choice and Purchase Quantity Price Sensitivities," *Marketing Science*, 7(1), 1-20

## References

- Krzanowski, Wojtek J. (1983), "Distance Between Populations Using Mixed Continuous and Categorical variables," *Biometrika*, 70 (1), 235-243.
- Kuhfeld, Warren F., Randall D. Tobias, and Mark Garratt (1994), "Efficient Experimental Design with Marketing Research Applications," *Journal of Marketing Research*, 31 (4), 545-557.
- Kullback, Solomon and R.A. Leibler (1951), "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22 (1), 79-86.
- Le, Nhu D., Weimin Sun, James V. Zidek (1997), "Bayesian Multivariate Spatial Interpolation with Data Missing by Design," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2), 501-510
- Lenk, Peter and Ambarg Rao (1990), "New Models from Old: Forecasting Product Adoption by Hierarchical Bayes Procedures," *Marketing Science*, 9, 42-53
- Lenk, Peter, Wayne S. Desarbo, Paul E. Green, and Martin R. Young (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, 15, 173-191
- Lenk, Peter and Michel Wedel (2001), "Bayesian econometrics: A Reaction to Geweke," *Journal of Econometrics*, 100, 79-80
- Liechty, John, Venkatram Ramaswamy, and Steven H. Cohen (2001), "Choice Menus for Mass Customization: An Experimental Approach for Analyzing Customer Demand with an Application to a Web-based Information Service," *Journal of Marketing Research*, 38 (2), 183-196
- Little, Roderick J. A. and Donald B. Rubin (1997), *Statistical Analysis with Missing Data*, New York: John Wiley & Sons
- Logman, Marc (1997), "Marketing Mix Customization and Customizability," *Business Horizons*, Vol.40 (6), 39-44

## References

- Lord, Frederic M. (1980), *Applications of Item Response Theory to Practical Testing Problems*, Hillsdale: Lawrence Erlbaum Associates.
- Manchanda, Puneet and Sunil Gupta (1997), "Complementarity in Shopping Baskets: Investigating Multi-category Purchase Incidence and Brand Choice," Working paper, Graduate School of Business, University of Chicago, Chicago, IL
- Manchanda, Puneet, Asim Ansari, and Sunil Gupta (1999), "The Shopping Basket: A Model for Multicategory Purchase Incidence Decisions," *Marketing Science*, 18(2), 95-114
- Marshall P. and Eric Bradlow (2002), "A Unified Approach to Conjoint Analysis Models," *Journal of American Statistical Association*, 97, 674-682
- Mehta, Raj and Eugene Sivadas (1995) "Comparing Response Rates and Response Content in Mail versus Electronic Mail Surveys," *Journal of the Market Research Society*, 37, 4, pp. 429-439
- Michalek, Jeremy J., Fred M. Feinberg, Feray Adiguzel, Peter Ebbes, and Panos Y. Papalambros (2005), "Realizable Product Line Optimization: Coordinating Product Positioning and Design for Heterogeneous Markets," *Under review Marketing Science*
- Montgomery, Alan L. (1997), "Creating Micro-marketing Pricing Strategies Using Supermarket Scanner Data," *Marketing Science*, 16, 315-337
- Montgomery, Alan and Peter E. Rossi (1999), "Estimating Price Elasticities with Theory-based Priors," *Journal of Marketing Research*, 36, 413-423.
- Montgomery, Alan L., Kartik Hosanagar, Ramayya Krishnan, and Karen B. Clay (2004), "Designing a Better Shopbot," *Management Science*, 50(2), 189-206
- Mulhern, Francis J. and Robert P. Leone (1991), "Implicit Price Bundling of Retail Products: a Multiproduct Approach to Maximizing Store Profitability," *Journal of Marketing*, 55, 63-76

## References

- Narasimhan, Chakravarthi, Scott A. Neslin, and Subrata Sen (1996), "Promotional Elasticities and Category Characteristics," *Journal of Marketing*, 60, 17-30
- Neff, Angela R. (1996), "Bayesian Two Stage Design Under Model Uncertainty," Ph.D Dissertation, Virginia Polytechnic Institute and State University
- Neelamegham, Ramya and Pradeep Chintagunta (1999), "A Bayesian Model to Forecast New Product Performance in Domestic and International Markets," *Marketing Science*, 18, 115-136
- Novak, Thomas P., Donna Hoffman, and Yiu-Fai Yung (2000), "Measuring the Customer Experience in Online Environments: A Structural Modeling Approach," *Marketing Science*, Vol. 19, No. 1, 22-42
- O'Hagan, Anthony (1994), *Kendall's Advanced Theory of Statistics, Volume 2B, Bayesian Inference*, New York: Wiley & Sons.
- Olkin, Ingram and R.F. Tate (1961), "Multivariate Correlation Models with Mixed Discrete and Continuous Variables," *Annals of Mathematical Statistics*, 32 (2), 448-465
- Orchard, T. and Max A. Woodbury (1972), "A Missing Information Principle: Theory and Applications," *Proc. 6<sup>th</sup> Berkeley Symposium on Math. Statist. and Prob.* 1, 697-715.
- Otter, Thomas, Sylvia Frühwirth-Schnatter, and Regina Tüchler (2003), "Unobserved Preference Changes in Conjoint Analysis," working paper, University of Vienna
- Popkowski, Peter T. L. and Ashish Sinha (2005), "A Methodology for Incorporating Prior Information into Choice Models," *Journal of Retailing and Consumer Services*, 12(2), 113-123

## References

- Press, James S. (2003), *Subjective and Objective Bayesian Statistics*, New Jersey: John Wiley and Sons
- Putler, Daniel S., Kirthi Kalyanam, and James S. Hodges (1996), "A Bayesian Approach for Estimating Target Market Potential with Limited Geodemographic Information," *Journal of Marketing Research*, 33, 134-149
- Raghavarao D., and Walter Federer (1979), "Block Total Response as an Alternative to the Randomized Response Method in Surveys," *Journal of the Royal Statistical Society, Series B*, 41, 40-45
- Raghu T. Santanam, Kannan P. K., Raghav H. Rao, and Andrew B. Whinston (2001), "Dynamic Profiling of Consumers for Customized Offerings Over the Internet: A Model and Analysis," *Decision Support Systems*, 32, 117-134
- Raghunathan, Trivellore and James Grizzle (1995), "A Split Questionnaire Survey Design," *Journal of the American Statistical Association*, 90, 54-63.
- Rahul, Telang, Peter Boatwright, and Tridas Mukhopadhyay (2004), "A Mixture Model for Internet Search Engine Visits," *Journal of Marketing Research*, 41(2), 206-214
- Rassler, Susanne (2002), *Statistical Matching, A Frequentist Theory, Practical Applications, and Alternative Approaches*, Springer.
- Rink, David R. (1987), "An Improved Preference Data Collection Method: Balanced Incomplete Block Designs," *Journal of the Academy of Marketing Science*, 1987, 15 (1), 54-57
- Rodgers, Willard L. (1984), "An Evaluation of Statistical Matching," *Journal of Business and Econometric Statistics*, 2, 91-102.
- Rossi, Peter E., Robert E. McCulloch, and Greg M. Allenby (1996), "The Value of Purchase History Data in Target Marketing," *Marketing Science*, 15(4), 321-340

## References

- Rossi, Peter E. and Andrew Ainslie (1998), "Similarities in Choice Behavior across Product Categories," *Marketing Science*, 17(2), 91-106
- Rossi, Peter E. and Greg M. Allenby (2003) "Bayesian Statistics and Marketing," *Marketing Science*, 22, 304-328
- Rossi, Peter E., Greg M. Allenby, and Robert E. McCulloch (2005), *Bayesian Statistics and Marketing*, New York: John Wiley and Sons
- Roszkowski, Michael J. and Andrew G. Bean (1990), "Believe It or Not! Longer Questionnaires Have Lower Response Rates," *Journal of Business and Psychology*, 4(4), 495-509
- Rubin, Donald B. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley&Sons.
- Rust, Roland T. and Peter C. Verhoef (2005), "Optimizing the Marketing Interventions Mix in Intermediate-Term CRM," *Marketing Science*, 24 (3), 477-489
- Sándor, Zsolt and Michel Wedel (2001), "Designing Conjoint Choice Experiments Using Managers' Prior Beliefs," *Journal of Marketing Research*, 38,430-449
- Savage, Leonard J. (1954), *The Foundations of Statistics*, New York: Wiley (reprinted by Dover, New York 1972)
- Schaffer, Joseph L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman&Hall
- Seetharaman, Seethu P. B., Andrew Ainslie, and Pradeep Chintagunta (1999), "Investigating Household State Dependence Effects Across Categories," *Journal of Marketing Research*, 36(November), 488-500
- Shoemaker, David M. (1973), *Principles and Procedures of Multiple Matrix Sampling*, Cambridge, MA: Ballinger.

## References

- Sikkel, Dirk and Adriaan W. Hoogendoorn (1995), "Models for Monthly Penetrations with Incomplete Panel Data," *Statistica Neerlandica*, 49 (3), 378-391.
- Sismeiro, Catarina and Randolph E. Bucklin (2004), "Modeling Purchase Behavior at an E-Commerce Web Site: Task Completion Approach," *Journal of Marketing Research*, 41(3), 306-323
- Smith, Adrian F. M. and Roberts G. O. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods (with discussion)," *Journal of the Royal Statistical Society, Series B*, 55, 3-23.
- Song Inseong, and Pradeep K. Chintagunta (2003), "A Discrete/Continuous Model for Multi-Category Behavior of Households," Working paper, University of Chicago
- Stremersch, Stefan, Allen M. Weiss, Benedict G. C. Dellaert, and Ruud T. Frambach (2003), "Buying Modular Systems in Technology-Intensive Markets," *Journal of Marketing Research*, 40(3), 335-350.
- Sudman, Seymour and Norman M. Bradburn (1989), *Asking Questions*. Oxford, Jossey-Bass.
- Talukdar, D., Sudhir K., and Andrew Ainslie (2002), "Investing New Production Diffusion across Products and Countries," *Marketing Science*, 21, 97-116
- Tanner, Martin A. and Wing H. Wong (1987), "The Calculation of Posterior Distributions by Data Augmentation (with discussion)," *Journal of the American Statistical Association*, 82, 528-550.
- Theil, Henri (1954), "Econometric Models and Welfare Maximization," *Weltwirtschaftliches Archiv*, 72, 19-70
- Ter Hofstede, Frenkel, Michel Wedel, and Jan-Benedict E. M. Steenkamp (2002), "Bayesian Prediction in Hybrid Conjoint Analysis," *Journal of Marketing Research*, 34, 253-261

## References

- Thayer, Dorothy T. (1983), "Maximum Likelihood Estimation of the Joint Covariance Matrix for Sections of Tests Given to Distinct Samples with Application to Test Equating," *Psychometrika*, 48 (2), 293-297.
- Toubia, Olivier, Duncan I. Simester, John R. Hauser, and Ely Dahan (2003), "Fast Poly-hedral Adaptive Conjoint Estimation," *Marketing Science*, 22, 3.
- Van der Linden, Wim J., and Jos J. Adema (1998) "Simultaneous Assembly of Multiple Test Forms," *Journal of Educational Measurement*, 35, 185-198
- Van der Linden, Wim J., Scrams D.J. , and Schnipke D.L. (1999), "Using Response Time Constraints to Control for Speediness in Computerized Adaptive Testing," *Applied Psychological Measurement*, 23, 195-210
- Van der Linden, Wim J. (2004), *Linear Models for Optimal Test Design*, New York: Springer Verlag
- Van der Linden, Wim J., Bernard P. Veldkamp, and James E. Carlson (2004), "Optimizing Balanced Incomplete Block Designs for Educational Testing," *Applied Psychological Measurement*, 28 (5), 317-331
- Van der Puttan, Peter, Joost Kok, and Amar Gupta (2002), Data Fusion through Statistical Matching, MIT Sloan School of Management, Working paper: 185
- Veldkamp, Bernard (2002), "Constrained Multidimensional Test Assembly," *Applied Psychological Measurement*, 26, 133-146
- Wedel, Michel and Rik Pieters (2000), "Eye Fixations on Advertisements and Memory for Brands: A Model and Findings," *Marketing Science*, 19, 297-312

## References

- Wedel, Michel and Peter Lenk (2001), "Bayesian Econometrics: A Reaction to Geweke," *Journal of Econometrics*, 100(1), 79-80
- Wind, Jerry and Arvind Rangaswamy (2001), "Customerization: the Next Revolution in Mass Customization," *Journal of Interactive Marketing*, 15(1), 13-32
- Wolfson, Lara J. (1995), "Elicitation of Priors and Utilities for Bayesian Analysis," Unpublished Ph.D. dissertation, Department of Statistics, Carnegie Mellon University
- Zhang, Junhua and Weiwu Fang (2003), "A New Approach of Information Discrepancy to Analysis of Questionnaire Data," *Communications in Statistics: Theory and Methods*, 32(2), 435-457
- Zhang, Jie and Michel Wedel (2004), "The Effectiveness of Customized Promotions in Online and Offline Stores," Working paper, University of Michigan Business School
- Zhang, Jie and Lakshman Krishnamurthi (2004), "Customizing Promotions in Online Stores," *Marketing Science*, 23(4), 561-578

## *References*

# Subject Index

- algorithm, 19, 20, 40, 41, 65, 67, 68, 72, 73, 77, 79, 84, 91, 108, 111, 117, 124, 126, 127, 128, 149, 158, 166, 168, 173, 174, 175
- Bayes, 16, 19, 21, 23, 24, 25, 26, 31, 74, 157, 166, 179, 181, 186, 207
- Bayes theorem, 23
- Bayesian analysis, 8
- Bayesian decision, 17, 18, 20, 21, 23, 30, 31, 33, 35, 122, 157, 165
- Bayesian Inference, 188
- Between block, 172
- Burn-in, 80, 118
- Convergence, 63, 118
- Correlation, 12, 19, 38, 48, 117, 125, 128, 129, 130, 131, 135, 136, 144, 145, 146, 158, 174
- Coupon, 15, 105, 107, 108, 122, 158
- covariance, 48, 57, 59, 62, 63, 70, 74, 94, 110, 115, 117, 119, 127, 131, 168, 173
- Customization, 1, 3, 4, 5, 6, 7, 10, 14, 15, 17, 18, 19, 22, 23, 104, 106, 107, 122, 157, 158, 163, 167, 173, 176, 180, 184, 186, 187, 193, 203
- Data augmentation, 26
- Data fusion, 44, 52, 53, 171
- Decision making, 7, 9, 17, 21, 22, 23, 29, 163
- Design matrix, 60, 66, 67, 73, 95
- Discriminant analysis, 13
- Estimation, 8, 9, 17, 18, 25, 27, 29, 30, 33, 35, 42, 46, 51, 62, 63, 97, 118, 125, 127, 159, 163, 165, 167
- Expenditure, 16, 21, 31, 108, 109, 111, 112, 113, 114, 115, 122, 124, 127, 128, 129, 130, 131, 132, 134, 135, 136, 143, 144, 145, 146, 147, 149, 156, 157, 158, 159, 166, 173
- Experimental design, 18, 20, 44, 46, 47, 58, 65, 91, 126, 165
- Factor analysis, 13
- Federov algorithm, 19, 20, 40, 41, 65, 68, 72, 73, 77, 79, 91, 108, 111, 124, 127, 149, 165
- Fisher information, 71
- Gamma distribution, 24
- General location model, 61, 94, 168
- Gibbs sampling, 19, 26, 69, 118, 119, 127, 128, 157
- Greedy algorithm, 91, 168, 174
- Heterogeneity, 17, 21, 26, 108, 110, 111, 116, 118, 127, 128, 129, 131, 165, 168
- Identification, 40, 56, 57, 62, 63, 64, 125, 172
- Imputation, 48, 63, 69, 71, 81, 84, 90, 168, 183
- Incidence, 17, 19, 109, 112, 115, 118, 127, 128, 129, 131, 135, 136, 143, 144, 145, 146, 147, 148, 158, 159, 166, 173, 174
- Incomplete block design, 44, 49, 55, 56, 57

## Subject Index

- Individual heterogeneity, 17, 127, 128, 129, 131, 168
- Information loss, 43, 47, 80, 91, 172
- Interdependence, 108, 110, 128, 129, 131
- Latent, 8, 21, 26, 113, 120, 165, 182
- Loglikelihood, 96
- Mahalanobis distance, 59, 61
- Marketing mix, 1, 3, 7, 14, 17, 131, 147, 164, 173, 174
- Markov Chain Monte Carlo, 25, 117, 184, 191
- Matrix sampling, 38, 41, 48, 72, 74, 79, 80
- Matrix sampling design, 38, 41, 48, 72, 74, 79, 80
- MCMC, 25, 28, 117, 118, 123, 125, 157, 164, 205
- Missing by design, 18, 40, 41, 43, 44, 47, 48, 54
- Missing information, 42, 71, 74, 77, 80, 81, 82, 83, 91, 97, 172
- Missing information principle, 71, 97
- Missingness, 43
- Mixed data, 61, 94, 168
- Multiple imputation, 42, 69, 71, 90, 168, 183
- Multivariate, 16, 19, 24, 26, 59, 61, 62, 64, 72, 94, 112, 117, 118, 119, 120, 121, 127, 157, 167, 168, 173, 180, 183, 186, 188, 190
- Noninformative priors, 70, 119
- Nonresponse, 11, 19, 97, 165
- Observed information, 96, 97
- Optimization, 19, 30, 34, 40, 59, 65, 91, 122, 123, 124, 149, 157, 163, 166, 167, 173, 176
- Posterior, 8, 23, 24, 25, 27, 28, 31, 32, 33, 34, 35, 69, 129
- Prior, 23, 28, 31, 33, 46, 92, 119
- Prior information, 1, 9, 13, 19, 28, 30, 33, 34, 35, 39, 46, 56, 62, 63, 64, 104, 108, 165, 168, 169, 170
- promotions, 14, 15, 22, 31, 32, 104, 105, 106, 107, 108, 109, 111, 123, 143, 148, 150, 157, 158, 159, 164, 167, 173
- Purchase panel, 50
- Questionnaire design, 12, 13, 18, 34, 39, 40, 41, 43, 46, 47, 52, 57, 58, 60, 61, 63, 64, 65, 66, 68, 69, 71, 72, 73, 74, 78, 79, 88, 90, 91, 92, 97, 104, 149, 166, 167, 168, 169, 171, 176, 206
- Simulation, 25, 42, 72, 74, 77, 127, 128, 129, 130, 131, 132, 133
- Split questionnaire, 10, 11, 12, 13, 18, 19, 23, 37, 38, 40, 41, 43, 44, 47, 49, 52, 53, 56, 58, 59, 60, 61, 62, 63, 64, 65, 66, 68, 69, 71, 72, 73, 74, 78, 79, 81, 84, 86, 87, 89, 90, 91, 92, 104, 149, 163, 165, 169, 170, 171, 176, 203, 205
- Subsampling, 44, 52

*Subject Index*

Sufficient statistics, 97

Time sampling, 38, 51

Tobit, 16, 19, 26, 112, 117, 118, 127,  
157, 166, 207

Within block, 42, 79, 82, 87, 168, 174

*Subject Index*

## Author Index

Adams L. L..	37	Chiang J.	172
Adiguzel F.	19	Chib S.	116, 143
Ainslie A.	108, 109, 115, 171	Chintagunta P.	170
<i>Allenby G.</i>	9, 21, 22, 23, 29, 32, 33, 34, 116	Chintagunta P.	108, 109, 171
Andrews R.	22	Choi J.	111
Anjos M. F.	121	Clark J. R.	37
Ansari A.	9, 22, 23, 104, 107, 114, 147	Cohen S.	116
Bachman J. G.	38	Comley P.	165
Bar-hen A.	59	Cook R.D.	125
Barnard J.	116	Cyert M. R.	30
Baumgartner R.	37, 178	Daudin J. J.	59
Bean A. G.	37	De Bruyn A.	166
Berdie D. R.	37	De Finetti B.	28
Berger J. O.	9, 24, 29, 30	DeGroot M. H.	30
Bertsimas D. J.	104	DeSarbo W. S.	111, 172
Blattberg R. C.	22	Dillman D. A.	37
Boatwright P.	22	Dorfman J. H.	34, 121
Bolton R.	108	Edwards Y. D.	116
Bradburn N. M.	69	Elrod T.	9
Bradlow E.T.	22, 23	Essagaier S.	104
Brockett P.L.	13	Fader P.	109
Bucklin R. E.	103, 111	Fang W.	13
Casella G.	24	Federer W.	57
Chaloner K.	29, 92	Federov V. V.	19, 20, 40, 41, 76, 125
		Fox E. J.	116, 117, 119

*Author Index*

Fram E. H.	165	Kohli R.	104
Frambach R. T.	4	Krishnamurthi L.	104, 111
Gale D.	37	Krzanowski W. J.	93
Garratt M.	40, 65	Kuhfeld W. F.	40, 65
Gelfand A. E.	25, 69	Kullback S.	40, 59
Gelman A.	164	Lattin J. M.	104
Geman S.	25	Le N. D.	44
Giesbrecht F.	55	Leibler R. A.	40, 59
Gilks W. R.	25	Lenk P.	9, 22, 23, 28, 29, 46
Gilula Z.	53, 63	Leone R. P.	109
Gooley C. G.	104	Liechty J.	22, 116
Grady D. B.	165	Little R. J. A.	40
Grizzle J.	12, 38, 42	Lodish L. M.	109, 116
Gupta S.	107, 108, 109, 114	Logman M.	3, 4
Haaland P.	13	Lord F. M.	45
Haldar S.	109	Manchanda P.	107, 109, 114
Heberline T. A.	37	Marshall P.	22
Hermkens P. L. J.	44	McCulloch R. E.	23, 63, 116
Herzog R. A.	38	Mehta R.	10
Hoffman D. L.	41, 77	Mela C.	104, 147
Hoogendoorn A. W.	38	Meng X.	116
Huber J.	46, 61	Mersereau A. J.	104
Jedidi K.	111	Michalek J.	22
Kalyanam K.	22	Montgomery A.	9, 22, 116
Kamakura W.	9, 53, 167	Mulhern F. J.	109
Kim J.	22	Nachtsheim C. J.	126
Kish L.	52	Narasimhan C.	109

*Author Index*

Neelamegram R.	22	Savage L. J.	28
Neff A. R.	48	Schaffer J.	71
Novak T. P.	41, 77	Seetharaman P. B.	108, 171
O' Hagan A.	61	Shoemaker D. M.	38, 41
Olkin I.	93, 164	Sikkel D.	38
Orchard T.	71, 96	Sinclair M.D.	37
Otter T.	22	Sinha A.	9, 29
Pieters R.	9	Sismeiro C.	23, 111
Popkowski P. T. L.	9, 29	Sivadas E.	10
Press S. J.	28, 31, 177	Smith A. F. M.	25, 69
Putler P. S.	9	Song I.	170
Raghavarao D.	57	Stremersch S.	4
Raghunathan T.	12, 38, 42	Sudman S.	69
Rahul T.	23	Sun W.	44
Raj S. P.	111	Talukdar D.	22
Ramaswamy V.	116	Tanner M.A.	26
Rangaswamy A.	1, 159	Tate R. F.	93
Rao A.	9	Ter Hofstede F.	9
Rassler S.	40, 42, 63, 168	Thayer D. T.	38
Rink D. R.	57	Theil H.	34
Roberts G. O.	25	Tobias R. D.	40, 65
Rodgers W. L.	63	van der Linden W. J.	46
Rossi P. E.	9, 21, 23, 32, 33, 34, 63, 104, 109, 115, 121, 171	van der Puttan P.	53
Roszkowski M. J.	37	Veldkamp B. P.	46
Rubin D. B.	40, 69, 180	Verhoef P. C.	1
Rust R. T.	1	Wedel M.	9, 19, 22, 23, 28, 29, 31, 46, 53, 92, 167, 186
Sandor Z.	9, 22, 29, 31, 46, 92	Wind J.	1, 159

*Author Index*

Wolfson L. J.	29
Wong W. H.	26
Woodbury M. A.	71, 96
Yung Y.	41, 77
Zhang J.	104
Zhang Ju.	13
Zidek J. V.	44
Zwerina K.	46

## Samenvatting (Summary in Dutch)

Het voornaamste doel van dit proefschrift is nieuwe methoden te ontwikkelen en te valideren om de effectiviteit van customization te kunnen bepalen en hoe dataverzameling kan worden verbeterd. Om deze doelstelling te realiseren, gebruiken we Bayesiaanse technieken in iedere fase van probleem oplossing, inferentie, schatten en besluitvorming. Dit proefschrift bevat twee essays voor twee verschillende problemen. Het eerste essay gaat over de vraag hoe dataverzameling kan worden verbeterd. Onze aanbevelingen zijn om gebruik te maken van gesplitste vragenlijsten, ook wel split questionnaires genoemd binnen de marketing, en we ontwikkelen een methodologie om tot optimale split questionnaires te komen. Het tweede essay betreft een promotie customization probleem waarbij meerdere productcategorieën tegelijkertijd worden beschouwd. We implementeren en schatten een model dat promotie-strategieën voor een combinatie van categorieën mogelijk maakt. De optimale strategie wordt verkregen via een combinatorische optimalisatie methode. We vatten hier de belangrijkste bevindingen samen.

### Samenvatting en Conclusies

In het eerste hoofdstuk hebben we de term "customerization" besproken. Marketing managers hebben behoefte aan online marketing-strategieën, zoals nieuwe methoden voor interactie met klanten, vanwege de verschuiving van marketing naar een online omgeving. Customization en customerization zijn twee verschillende concepten die niet met elkaar

moeten worden verward. Customerization geeft aan dat interesses en desinteresses van de individuele klant centraal staan in iedere fase van het marketing-proces, in plaats van dat alleen het aanbod van het product op de klant wordt toegespitst. Customerization kan kort worden omschreven als het combineren van strategieën binnen 1-op-1 marketing, personalization, targeting en mass-customization. Succesvolle customerization strategieën combineren de vraag- en aanbodzijden van de markt. Met het oog hierop worden de mogelijkheden van het afstellen van marketing-mix instrumenten voor kopers en verkopers besproken in Hoofdstuk 1. Om nieuwe customerization strategieën te kunnen ontwikkelen, is informatie over de klant benodigd (interesses en desinteresses, levensstijl, aankoopgedrag, achtergrond variabelen, etc.). Deze informatie is onmisbaar voor het identificeren, differentiëren en interacteren met klanten, en kan worden verzameld met behulp van vragenlijsten. Betere methoden om data te verzamelen en betere methoden voor customization zullen managers en marktonderzoekers kunnen helpen om betere beslissingen te nemen. De Bayesiaanse gereedschapskist zal hierbij zeer nuttig blijken. We geven twee toepassingen waarin we data verzamelen met betrekking tot zogenaamde "zachte" variabelen, zoals levensstijl of klanttevredenheid in Hoofdstuk 3 en promoties over meerdere categorieën in Hoofdstuk 4.

In het tweede hoofdstuk gaan we dieper in op de vraag waarom de Bayesiaanse methodologie bij uitstek geschikt is voor de besluitvormingsoriëntatie bij marketing-problemen. Vanuit Bayesiaans

### *Samenvatting (Summary in Dutch)*

oogpunt wordt alle informatie gebruikt om de mate van onzekerheid in inferentiële en beslissingsproblemen zoveel mogelijk te reduceren. Een belangrijke reden voor het toegenomen gebruik van Bayesiaanse methoden binnen marketing in de afgelopen tien jaar is niet alleen de toegenomen capaciteit van computers en het succes van MCMC algoritmen om complexe problemen op te lossen, maar ook de karakteristieken van marketing-data, de noodzaak om marketing-problemen als een beslissingsprobleem te benaderen, en de flexibiliteit en robuustheid van Bayesiaanse methoden. Zonder de intentie te hebben alle mogelijkheden op te noemen, stellen we dat marketing-modellen met latente variabelen, ontbrekende data, gemengde verdelingen, heterogeniteit in coëfficiënten, niet-lineariteiten en discrete data gemakkelijk kunnen worden geschat binnen het Bayesiaanse raamwerk. Omdat het Bayesiaanse paradigma alle informatie gebruikt en prior informatie combineert met waargenomen data om modellen te schatten (informatie wordt bijgesteld), is het Bayesiaanse paradigma optimaal voor beslissingsproblemen binnen marketing. Het Bayesiaanse besluitvormingsproces houdt rekening met onzekerheid intrinsiek, aanwezig in zowel het model als in het schatten van de parameters.

In het derde hoofdstuk leggen we ons toe op split questionnaires, in plaats van de gangbare lange vragenlijsten binnen marketing (die doorgaans meer dan 20 minuten in beslag nemen). Split questionnaires hebben het potentieel om hoge-kwaliteit informatie sneller van respondenten te verkrijgen, en tegen aanzienlijk lagere kosten. In split questionnaires beantwoorden verschillende respondenten verschillende delen van de vragenlijst. Dit betekent dat we verschillende versies van de

### *Samenvatting (Summary in Dutch)*

vragenlijst hebben die korter zijn dan de volledige vragenlijst. Na verschillende split questionnaires te hebben gegenereerd en deze aan respondenten te hebben toegewezen, imputeren we de data voor de ontbrekende onderdelen door de antwoorden van andere respondenten te gebruiken die deze onderdelen van de vragenlijst wel hebben beantwoord. Op het eind hebben we bijna dezelfde hoeveelheid informatie via de split questionnaire als via de uitgebreide volledige vragenlijst, maar in minder tijd, tegen lagere kosten and met antwoorden die van betere kwaliteit zijn (minder item non-response, hogere response en nauwkeurigere antwoorden). We suggereren een methodologie om versies van split questionnaires te genereren, die zijn gebaseerd op bepaalde prior informatie en waarvoor we optimale experimentele design methoden gebruiken. We gebruiken de Kullback-Leibler afstand als een criterium, genereren designs met het Modified Federov algoritme om over de gehele design-ruimte te zoeken, en illustreren dat goede designs gevonden kunnen worden. We gebruiken synthetische data om de prestaties van het algoritme te illustreren, echte data om de statistische efficiëntie te illustreren (m.a.w. we tonen aan dat we bijna dezelfde informatie verkrijgen met split questionnaires als met volledige vragenlijsten), en een veldstudie geeft efficiëntie in termen van het gedrag van respondenten aan. Statistische en gedragsmatige efficiëntie van de split questionnaire designs worden vastgesteld door deze designs te vergelijken met volledige vragenlijsten of met kleinere vragenlijsten geconstrueerd op basis van ad-hoc methoden.

### *Samenvatting (Summary in Dutch)*

In het vierde hoofdstuk wordt het promotie-design probleem besproken waarbij meerdere productcategorieën tegelijkertijd worden beschouwd. Tot nu toe beperken veel modellen waarin meerdere categorieën worden beschouwd zich tot slechts een klein aantal categorieën, en is het voornaamste doel van deze studies om elk type verband tussen de vraag naar verschillende productcategorieën te begrijpen (substitutie, complementariteit of onafhankelijkheid). Retailers kunnen deze verbanden tussen verschillende categorieën gebruiken voor moment-van-aankoop materiaal, coupons voor meerdere categorieën, voor creatieve winkelinrichting, en voor online advertenties. Binnen marketing hebben we echter betere modellen nodig om de preferenties van consumenten met betrekking tot categorieën te begrijpen. In onze toepassing beschouwen we niet alleen verbanden tussen categorieën, maar ook de verbanden met betrekking tot beslissingen om wel-of-niet tot aankoop over te gaan en het uit te geven bedrag. Dit alles is geformaliseerd in een hierarchisch Bayes type-2 tobit model. We benaderen het probleem van het optimaal promoten van een beperkte set van categorieën als een combinatorisch optimalisatieprobleem en we gebruiken een algoritme om vele promotie-designs te genereren en de beste te vinden. De doelstellingsfunctie is de maximale verandering in winst wanneer de categorie wordt gepromoot voor de geselecteerde categorieën. Onze methode kan ook worden toegepast op andere mogelijke marketing-aanbiedingen, die betrekking kunnen hebben op een product, een dienst, een combinatie hiervan, of op een bundel van producten en/of diensten. Om het model te schatten en te optimaliseren gebruiken we Bayesiaanse technieken, die het mogelijk maken onzekerheid mee te nemen. Voor beslissingsproblemen binnen de

*Samenvatting (Summary in Dutch)*

marketing is het belangrijk om de mate van onzekerheid in het nemen van beslissingen effectief te communiceren naar managers. Analyses waarin exogene variabelen worden meegenomen kunnen voorspellingen genereren die sterk in precisie kunnen variëren. Het is dus inderdaad van groot belang dat onzekerheid in voorspellingen wordt gecommuniceerd naar managers, zodat zij de voorkeur kunnen geven aan beslissingen die minder onzekerheid met zich meebrengen, of zodat zij meer informatie kunnen verzamelen om de mate van onzekerheid verder terug te dringen.