Market Segmentation Using Bayesian Model Based Clustering

ISBN: 978-90-393-51253
Printed by: Gildeprint B.V., Enschede

Market Segmentation Using Bayesian Model Based Clustering

Marktsegmentatie met behulp van Bayesiaanse Modelgebaseerde Clustering

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. J.C. Stoof,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 20 november 2009 des middags te 4.15 uur

door

Pascal van Hattum

geboren op 22 oktober 1976, te Amersfoort

Promotor: Prof. dr. H. J. A. Hoijtink

Acknowledgements

This dissertation is one of the results of the Ph.D project I have been doing in the past few years. The goal of this project was to bring scientific work into business perspective and vice versa. During the project I was working both at the Department of Methodology and Statistics and The SmartAgent Company. Through this combination I was able to work with all kind of experts from both worlds. The work I have done could not have been accomplished without the cooperation, support and motivation of these experts and a lot of other persons. I owe my gratitude to all these people. However, I owe my special thanks to a few people and would like to mention them briefly:

Herbert Hoijtink: First and foremost I offer my sincerest gratitude to you. Your always enthusiastic and inspiring supervision with your down-to-earth attitude are an example for me. I have learned a lot from you, both scientifically and personally. It has been a pleasure working with you and I hope that this will continue in the future.

Douwe Reitsma: Thanks for giving me the opportunity to work part-time on this dissertation and supporting me from The SmartAgent Company. I appreciate your always uncensured comments about my work. It has been very helpful for me to put this work into business perspective.

Irene Klugkist: I remember our first meeting about doing a part-time Ph.D project. Your enthusiasm was one of the reasons to start this project. I have never regretted this decision.

My colleagues from The SmartAgent Company and the Department Methodology and Statistics: Two types of colleagues from which I have learned a lot. I enjoy working with you.

My friends: Chatting, e-mailing, drinking, eating, discussing, laughing, crying, sporting, diving you are there for me to relax and to keep my feet on the ground.

Richard Cornelissen: Thanks for designing the cover and the invitation of this dissertation.

Marcel van Hattum: Thanks for just being my brother.

My parents: Thanks for your everlasting support, proud and interest in what I do.

Marieke Ruijzing: Although astonished that this dissertation was a hobby for me and 'nerdje spelen' is not very stimulating, you were always there for me to support me, to be proud in what I do and to show me that there are more important things than work. Thanks for your love and being the mother of my little baby.

A little Van Hattum: A new life in mom's belly. With your presence I am assured of a new challenge in life.

Contents

Chapter 1

Introduction

This dissertation deals with two basic problems in marketing, that are market segmentation, which is the grouping of persons who share common aspects and market targeting, which is focussing your marketing efforts on one or more attractive market segments.

In order to conceptualize market segmentation, imagine flying in a hot air balloon high in the sky. When you look at the people below, they appear remarkably similar. If the balloon descends, some differences become clear: you can see short and tall people, slim and fat. If the balloon descends to street level and you join these persons on the street, you can discover that each person is in some respect unique, but that there are also similarities that you did not see before. You can see men and women, people who are well-dressed and people who are more casually dressed. Some of them are obviously happy, others are not. If you speak to these persons you can discover even more similarities. Some persons have a more adventurous attitude towards life, whereas others are more sober. Some persons have a tendency to drive luxury cars, whereas other persons do not care about cars at all. It seems that all these people are at the same time similar and different. This balloon ride resembles the concept of market segmentation. First you look at a crowd of people as a whole. Then you study them from a closer point and discover the characteristics in which these persons can be differentiated. Finally, you look at those groups of persons who share common characteristics.

Most statisticians do not have the privilege to do a segmentation study from a hot air balloon and are designated to data sets. These data sets are a simplification of reality and contain characteristics, often gathered using market research, from the research population under study. The goal is to find groups of persons who share common characteristics in these data sets. To give a simple example, when we ask a group of respondents (that are persons who participated in a market research study) to pick at least one item from the following list of statements:

1. it's important to have the latest car model,
2. it's important to have a lot of features on my car,
3. the price of the car must be as cheap as possible,
4. the main reason for having a car is not getting wet,

it is most likely that we may find two market segments or clusters of persons with more or less the same attitude towards cars. Cluster 1 containing respondents who have a higher probability of picking statement 1 and 2, and a lower probability of picking statement 3 and 4. For persons from this cluster a luxury car is a must have, they want a luxury drive. And Cluster 2 containing respondents who have a higher probability of picking statement 3 and 4, and a lower probability of picking item 1 and 2. For persons from this cluster a car is just a 'thing' to bring you from A to B.

For this simple example it is clear what kind of clusters can be expected. However in marketing the data sets under study often contain a large number of items and a large number of respondents. In large data sets it is not obvious how many clusters there are and how the clusters can be identified. In order to find the clusters in these large data sets, one needs statistical clustering techniques. In fact, much of the literature about market segmentation has evolved around the techniques of identifying clusters from data. Substantial parts of this literature are comparative papers that contrast the most widely used clustering techniques (MacLachlan and Mulhern, 2004). More recent papers (MacLachlan and Mulhern, 2004; Magidson and Vermunt, 2002; Mulhern and MacLachlan, 2003) compare model based algorithms with more traditional cluster techniques, like K-means.

Within the context of segmentation a number of papers do suggest better segmentation results when model based clustering algorithms are used (MacLachlan and Mulhern, 2004). An important advantage of model based clustering (Bensmail et al.,1997; Fraley and Raftery, 1998; Vermunt and Magidson, 2000, p. 1-2, 152) over traditional clustering techniques (Hair et al., 1984, p. 469-518) is the statistical framework model based clustering is based on. A disadvantage is that model based clustering approaches are less available in popular statistical software than more traditional statistical models. This results in researchers making their own software, like for example: Glimmix (Wedel and Kamakura, 2000, p.181-186) and LatentGold (Vermunt and Magidson, 2000).

In Chapter 2, *'Market Segmentation using Brand Strategy Research: Bayesian Inference with respect to Mixtures of Log-Linear Models'*, published as Van Hattum and Hoijtink (2009b), a Bayesian model based clustering approach for dichotomous item responses is presented. In this chapter it is shown how the clustering approach can deal with missing values, large data sets and within cluster item dependencies. Furthermore, the consequences of using a cluster model assuming local within cluster item independence when in the data there are local within cluster item dependencies will be illustrated. The examples in this chapter are illustrated by using Brand Strategy Research, that is a theoretical framework to make motivational groups or clusters.

Although the clustering approach in Chapter 2 gives the statistically optimal number of clusters, this solution often (especially in the case of large data sets) contains too many clusters for the intended marketing purposes. From this it is clear that market segmentation is not solely a matter of statistics, but an interaction between statistics and marketing. Or, as MacLachlan and Mulhern (2004) describe this perfectly: *'in any empirical problem, the researcher must necessarily use a substantial dose of subjectivity and domain knowledge. This can be aided by computation of some statistical indicators, but ultimately the decision, regarding the number of clusters to use in any particular problem, will be the result of viewing those indicators in the light of the marketing decision problem at hand'.* In Chapter 3, *'Reducing the Optimal to a Useful Number of Clusters for Model Based Clustering'*, six criteria of good market segmentation, an information criterion and two conjectures describing the geometry of model based clustering models are used to reduce the statistically optimal number of clusters to a smaller number, suited for the intended marketing purposes.

As mentioned before, substantial parts of the literature about the techniques of identifying clusters from data contrast the most widely used clustering techniques. However, papers that actually compare different model based clustering approaches are scarce (Meila and Heckerman, 2001; Ter Braak et al., 2003). In Chapter 4, *'A Comparison of Model Based Clustering Algorithms'*, a comparison is made between the Bayesian clustering approach described in Chapter 2 and the approaches implemented in LatentGold and Glimmix. Using simulation studies the performance of the approaches is evaluated.

In the previous three chapters the emphasize is on the techniques of identifying market segments or clusters in data. Once the market segments or clusters are identified, marketeers evaluate the attractiveness of each cluster. The next step in marketing is targeting, which is described in the last two chapters. Market targeting is focussing your marketing efforts on the most profitable clusters. One such focus is differentiated marketing. Or, in other words, marketeers may be trying to sell exactly the same product or service, but it will change, for example, its promotional methods for each cluster.

Differentiated marketing as a market targeting strategy can also conceptualized using the hot air balloon ride and the car market example. From the balloon ride we learned that persons on the street are at the same time similar and different. For example, we learned that many persons drive cars, but the attitude towards cars may be different among clusters of persons. Imagine yourself as a marketeer with the mission to sell cars. Using the knowledge of your balloon ride you can develop different promotions in order to sell the same product, that is cars. Each promotion with a different tone-of-voice. A tone-of-voice that is suitable for the cluster to be targeted. For example, in a promotional campaign for a cluster that mainly contains families with young children, the emphasize of the promotional text must be on the safety of the car. Or, the space of the car. For a cluster that mainly contains persons who are looking for luxury, the emphasize of the text must be on the luxury components of the cars. Or, the number of horse powers. Different attitudes towards cars must lead to different promotional campaigns.

In order to target clusters as individually as possible, it is important to learn a lot about the persons in these clusters. It seems that collecting all the desired customer information in one single source questionnaire is the best solution. But as time and money is limited in most marketing companies, this is often not realized. An attractive and practical solution is data fusion, or, in other words, integrating different data sets.

In Chapter 5, *'The Proof of the Pudding is in the Eating. Data Fusion: An Application in Marketing'*, published as Van Hattum and Hoijtink (2008) and Chapter 6, *'Improving your Sales with Data Fusion'*, published as Van Hattum and Hoijtink (2009a) it is shown how the results of two market segmentation studies are fused to two customer databases. To select the best data fusion algorithm, two traditional data fusion methods, that are polytomeous logistic regression and a nearest neighbor algorithm, are compared with two model based clustering approaches. Using the fused data sets, cluster specific questionnaires and cluster specific catalogues are made and send out to the customers. The effectiveness and profitability of each data fusion algorithm are determined using internal and external criteria.

The main goal of the research in this dissertation was to bring scientific work about segmentation and targeting into business perspective. To be as realistic as possible most of the data sets used are from marketing businesses. All the research in this dissertation has successfully been tested and used in day-to-day business of The SmartAgent Company.

Chapter 2

Market Segmentation Using Brand Strategy Research: Bayesian Inference with respect to Mixtures of Log-Linear Models

Abstract[*]

This chapter presents a Bayesian model based clustering approach for dichotomous item responses that deals with issues often encountered in model based clustering, like missing values, large data sets and within cluster dependencies. The approach proposed will be illustrated using an example concerning Brand Strategy Research.

[*] This chapter is accepted for publication as Van Hattum, P. and Hoijtink H. (2009b). Market Segmentation using Brand Strategy Research: Bayesian Inference with respect to Mixtures of Log-Linear Models. *Journal of Classification, 16.*

2.1 Introduction

The popularity of model based clustering has increased in recent years (Ter Braak et al., 2003; Vermunt and Magidson, 2000; Wedel and Kamakura, 2000, p.75). LatentGold (Vermunt and Magidson, 2000) and Glimmix (Wedel and Kamakura, 2000, p.181-186) are the two best known software packages that can be used for the clustering of dichotomous item responses assuming that the responses of persons within clusters are independent. Both packages share a number of features:

- Both use the EM-algorithm (Dempster et al., 1977) to obtain parameters estimates.

- Both require the specification of a range for the number of clusters. For each number of clusters an information criterion is computed, that can be used to select the best number of clusters.

- Both offer available case analysis assuming that the missing values are missing at random (MAR) (Schafer and Graham, 2002) have the possibility to analyze the observed data.

Also assuming within cluster independence, the approach described in this chapter uses a hierarchical algorithm (Hoijtink and Notenboom, 2004) to simultaneously determine the number of clusters and obtain parameter estimates. Note that a prespecified range for the number of clusters is not necessary. An estimate of the number of clusters is an outcome of the hierarchical algorithm used. The algorithm used can handle missing values assuming that they are missing at random (MAR). Furthermore, both the distribution of the data under the null-hypothesis and a user specified missing value mechanism are used to determine the null-distribution of a goodness-of-fit statistic such that the missing values are accounted for.

However, the main contribution of the approach presented in this chapter is that it can handle cluster specific interactions among the items. This model was described in Hagenaars (1988). This model can not be handled by LatentGold (which can handle interactions, but not cluster specific interactions) and Glimmix (a generalized linear model is a special case of a log-linear model in which the interaction terms among the predictors are excluded). A further contribution is that large data sets can easily be handled with the Bayesian computational approach described in this chapter. To give one benchmark, a simulated data set assuming within cluster independence consisting of 156 items, 50,000 persons and 13 clusters, was analyzed in 15 hours using the approach proposed and over 60 hours using

LatentGold[†]. With Glimmix it was not possible to analyze the simulated data set[‡].

The structure of this chapter is as follows. Section 2.2 describes the data set that is used in this chapter. Furthermore, this section describes the underlying theory of Brand Strategy Research (BSR), which is a framework to make motivational groups or clusters. This section also introduces prior knowledge about the underlying theoretical framework of BSR. In order to use this prior knowledge we propose a Bayesian model based cluster model. Section 2.3 introduces the proposed cluster model, which is a latent mixture of log-linear models, that is able to deal with the prior knowledge. This section also shows how the model parameters are estimated using a Bayesian computational framework. Section 2.4 deals with a difficult question commonly asked in cluster analysis: 'how to determine the number of clusters?'. In Section 2.5 the results of applying the clustering algorithm on the data set at hand are given. The consequences of using a cluster model assuming local item independence within the latent clusters when there are dependencies among the items within the latent clusters will be illustrated. This section will also describe the validation of the theory of Brand Strategy Research. This chapter concludes with a short discussion in Section 2.6.

2.2 Description of the Data Set Used

The data that are used in this chapter come from *Brand Strategy Research (BSR)* (Brethouwer et al., 1995, p. 8; Oppenhuisen, 2000, p. 79-81), which is a methodology of making motivational groups or clusters. BSR is based on Adler's social-psychology theory (Callebaut et al., 1999, p. 55-60) and provides a framework for understanding customers at the 'deepest' level. This motivational level gives knowledge of consumer's fears, beliefs and values, thus providing an understanding of the fundamental motivations that drive (future) purchase decisions of customers. The BSR framework consists of a strategic map in which all BSR information (that is the content of the BSR questionnaire, which is described below) is presented. Two axes divide the map. The first (horizontal) axis is called the 'sociological' axis and indicates how a person relates to their social environment: the right side indicates involvement (belonging), the left side indicates independent (affirmation). The second (vertical) axis is called the 'psychological' axis and indicates how a person handles with 'tensions': the top side indicates an expression of 'tensions' (extravert) and the bottom side indicates a suppression or ignorance of 'tensions' (introvert). The result is a four-quadrant strategic map as shown in Figure 2.1.

[†] In order to speed up LatentGold the bi-variate residuals were not calculated.

[‡] Glimmix allows you to analyze up to 150 variables and up to 50,000 persons.

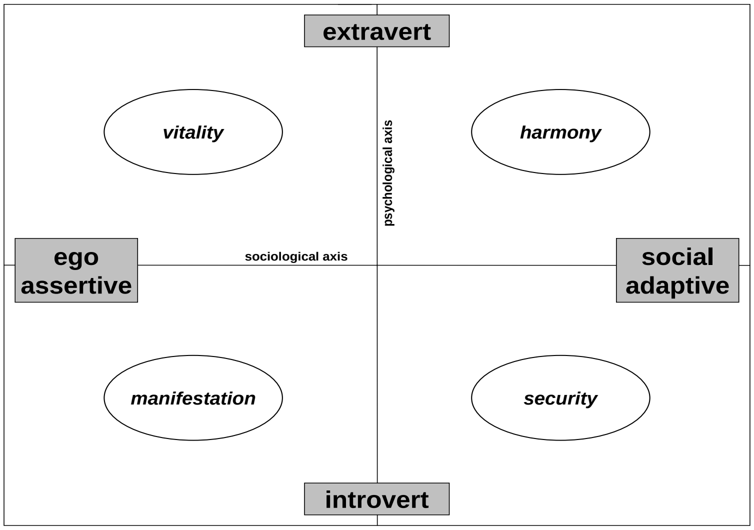


Figure 2.1: BSR Strategic Map

The idea behind BSR is that the four quadrants in the strategic map represent four main motivational clusters which can be found in each researched domain. Each of these clusters demonstrates unique needs, motivations and products or services and communication requirements. In a researched domain it is also possible that mixtures of these four main clusters are found. The four main motivational clusters are:

1. In the upper left quadrant a cluster that is described with the word ’vitality’. Persons from this cluster are self conscious, self-confident in their attitude towards (choices in) life and energetic, vital and passionate in their behavior.

2. In the lower left quadrant a cluster that is described with the word ’manifestation’. Persons from this cluster are career oriented and aspire a certain (high) status in life in connection with certain status symbols and conspicuous consumption.

3. In the upper right quadrant a cluster that is described with the word ’harmony’. Persons from this cluster strive for harmony in every aspect of life and harmonious relations with all people they meet in daily life.

4. In the lower right quadrant a cluster that is described with the word 'security'. Persons from this cluster are mainly oriented on their peer group and the rules and values of this group.

The whole BSR questionnaire consists of five questions, each containing multiple psychographic items. The first question contains items that describe a person's character. The second question tells something about a person's type of household. The third gives a person's occupations, the fourth question tells something about a person's hobbies and interests and the last questions tells which values a person can have in live. Appendix I displays the BSR questionnaire. In total there are 149 psychographic items answered by 2294 respondents. For each question a respondent has to pick the items which describe the person he has in mind the best. Because each question contains a broad range of items it is unlikely that a respondent can not pick an item from the item list. If a respondent did not pick any of the items from an item list, we assume that the respondent skipped the whole question accidently. The answers to the items from these skipped questions are considered to be missing values.

As described above the content of the BSR questionnaire can be presented in the strategic map as in Figure 2.1. From past experience it can be said that this content is presented in more or less the same way in the strategic map in each researched domain (Brethouwer et al., 1995, p. 8; Oppenhuisen, 2000, p. 79-81). For example, persons who are assigned to the main motivational cluster that can be described by the word 'Vitality' are more likely to pick items like for example: 'Adventurous' (character traits), 'Single' (household types), 'Entrepreneur' (occupations), 'Snow boarding' (hobbies) and 'Independence' (values). And persons who are assigned to the main motivational cluster that can be described by the word 'Manifestation' are more likely to pick items like for example: 'Self-assured' (character traits), 'Busy dynamical family' (household types), 'Manager' (occupations), 'Build a successful career' (hobbies) and 'Success in life' (values). Within each motivational cluster, not only the individual items are more likely to be picked, but also pairs of items are more likely to be picked. For example, in the case of the cluster that can be described by the word 'Vitality', persons who pick item 'Adventurous' (character traits) are more likely to pick also 'Snow boarding' (hobbies). And in the case of the cluster that can be described by the word 'Manifestation', persons who pick item 'Manager' (occupations) are more likely to pick also item 'Success in life' (values). Likewise all other items from the BSR questionnaire can be pre-assigned to one of the four main motivational clusters. As a matter of fact, in the past years the BSR questionnaire has been subjected to research to fill up the four main motivational clusters. This results in a substantial dose of prior knowledge about the underlying theoretical framework. This prior knowledge is used in Section 2.5 to determine which interaction effects are included in our model in order to find the four main motivational clusters. In other words, from past experience it is known which combinations of items are more likely to be picked and in Section 2.5 this prior knowledge is used in our model based clustering approach.

As will be elaborated in the next section, within each cluster the data will be modelled with a log-linear model. In large data sets it is impossible to account for all dependencies among the items. In the example data set there are 149 items. This would lead to 149 main effects, 11026 two-way interaction effect, et cetera, for each cluster. With such a huge model the data will be over-fitted. The result will probably be a one or two cluster model for which an interpretation may be hard or impossible to find.

To keep the size of the model under control, in this chapter we will only consider main effects and pre-specified sets of two-way interactions within each cluster. As indicated above (and further elaborated in Section 2.5), knowledge with respect to the research domain can be used to specify the most important interactions. As will be illustrated in Section 2.5, if a model without two-way interactions is used, an interpretation has to be found for a solution consisting of 35 clusters. If a pre-specified set of two-way interactions is used, the solution consists of 5 clusters for which an interpretation is straightforwardly obtained.

The empty set (that is a model without two-way interactions) may render a solution with many clusters that has a nice fit. Different sets of two-way interactions may lead to different solutions that also have a nice fit. Which of these solutions is the best, is a question that may not have an answer. Perhaps the question to ask is not which solution is the best, but which solution has an acceptable fit and an interpretation that is useful for the research question at hand. As will be explained in Section 2.4 and illustrated in Section 2.5, if the goodness of fit of a cluster solution is not sufficient, an inspection of residuals can be used to determine if and which two-way interactions should be added to the model under consideration.

2.3 Model Based Clustering Algorithm

2.3.1 Introduction

An important difference between standard clustering (Hair et al., 1984, p. 469-518) and model based clustering (Banfield and Raftery, 1993; Bensmail et al., 1997; Fraley and Raftery, 1998; Newcomb, 1886; Pearson, 1894; Vermunt and Magidson, 2000, p. 1-2, 152) is that in the latter it is assumed that the data are generated by a certain mixture of underlying probability distributions. An advantage of this probabilistic approach is that the cluster criterion (Hair et al., 1984, p. 482-490; Wedel and Kamakura 2000, p. 39-73), which is usually difficult to define and calculate for complex models, is not needed.

The probabilistic approach is used to develop a Bayesian model based clustering approach that deals with missing values, large data sets and within cluster dependencies. In Section 2.3.2 the probability densities used in our model based clustering algorithm are described. This section also shows how the proposed algorithm accommodates local within cluster dependencies. In Section 2.3.3 the core of the model based clustering algorithm, that is the Gibbs sampling algorithm, is explained. This section shows how the Bayesian framework breaks down a rather complex problem into smaller subproblems in order to get a sample from the posterior distribution. This section also shows how data augmentation is used to handle the missing values. Section 2.3.4 briefly describes the hierarchical algorithm developed by Hoijtink and Notenboom (2004). A small simulation study is included to illustrate this algorithm.

2.3.2 Likelihood, Prior Distribution and Posterior Distribution

Let x_{ij} denotes the response of respondent $i = 1, \ldots, N$ to item $j = 1, \ldots, J$, $x_{ij} \in \{0, 1\}$, where 1 indicates that respondent i picked item j and 0 indicates that respondent i did not pick item j. The $N \times J$ matrix $\boldsymbol{X}$ contains the item responses. The J vector $\boldsymbol{x}_i$ is defined as a vector containing the response pattern or item responses of respondent i. The N vector $\boldsymbol{x}_j$ is defined as a vector containing the responses of the respondents to item j. The data matrix $\boldsymbol{X}$ is split into a data matrix $\boldsymbol{X}^0$ and $\boldsymbol{X}^*$, that is $\boldsymbol{X} = \{\boldsymbol{X}^0, \boldsymbol{X}^*\}$ and $J = J^0 + J^*$. The part $\boldsymbol{X}^0$ is the $N \times J^0$ data matrix containing J^0 items that, within the latent clusters, are independent of the other items. The part $\boldsymbol{X}^*$ is the $N \times J^*$ data matrix containing the J^* items that, within the latent clusters, are dependent on some of the other items. The part $\boldsymbol{X}^*$ is split into K subsets. Within each latent cluster there is within subset dependence among the items and between subset independence among the items. Let $\boldsymbol{X}^k$, for $k = 1, \ldots, K$, be the $N \times J^k$ data matrix containing the k^{th} set of J^k locally dependent items, that is $\boldsymbol{X}^* = \{\boldsymbol{X}^1, \ldots, \boldsymbol{X}^K\}$ and $J^* = \sum_{k=1}^{K} J^k$. Similarly, $\boldsymbol{x}_i = \{\boldsymbol{x}_i^0, \boldsymbol{x}_i^*\}$, where $\boldsymbol{x}_i^* = \{\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^K\}$. Note finally that the matrix $\boldsymbol{M}$ is a $N \times J$ indicator matrix with elements m_{ij}, where a 1 indicates that a response is missing and a 0 that a response is observed.

Within each cluster each item of the J^0 locally independent items is characterized by a parameter $\pi_{j|q}$, that is the probability of responding 1 to item j in cluster q. Note that $\boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_q, \ldots, \boldsymbol{\pi}_Q\}$ and $\boldsymbol{\pi}_q = \{\pi_{1|q}, \ldots, \pi_{j|q}, \ldots, \pi_{J^0|q}\}$.

Within each cluster a log-linear model containing main effects and two-way interaction effects is used to model the responses to $\boldsymbol{X}^k$, for $k = 1, \ldots, K$. The parameters of these log-linear models are denoted by $\boldsymbol{\lambda}_q^k$. A further elaboration of these log-linear models follows below.

Let $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_q, \ldots, \omega_Q\}$ be the Q vector containing the cluster weights, that is the proportion of persons allocated to each cluster. Finally, let $\boldsymbol{\tau}$ be the N vector containing the unobserved cluster memberships for each person $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_i, \ldots, \tau_N\}$, where $\tau_i \in \{1, \ldots, Q\}$.

The complete data likelihood of the model based cluster model is given by:

$$L(\boldsymbol{X} \mid \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\omega}) = \prod_{i=1}^{N} \sum_{q=1}^{Q} P(\boldsymbol{x}_i \mid \tau_i = q)\omega_q, \tag{2.1}$$

where

$$P(\boldsymbol{x}_i \mid \tau_i = q) = P(\boldsymbol{x}_i^0 \mid \tau_i = q) P(\boldsymbol{x}_i^* \mid \tau_i = q). \tag{2.2}$$

As is elaborated in the next subsection, missing values are dealt using data augmentation. For the locally independent items,

$$P(\boldsymbol{x}_i^0 \mid \tau_i = q) = \prod_{j=1}^{J^0} \pi_{j|q}^{x_{ij}} (1 - \pi_{j|q})^{1 - x_{ij}}. \tag{2.3}$$

For the locally dependent items,

$$P(\boldsymbol{x}_i^* \mid \tau_i = q) = \prod_{k=1}^{K} P(\boldsymbol{x}_i^k \mid \tau_i = q). \tag{2.4}$$

The number of possible response vectors in subset k is 2^{J^k} and is denoted by $\boldsymbol{Y}^k = \{\boldsymbol{y}_1^k, \ldots, \boldsymbol{y}_p^k, \ldots, \boldsymbol{y}_{2^{J^k}}^k\}$ and $\boldsymbol{y}_p^k = \{y_{p1}^k, \ldots, y_{pj}^k, \ldots, y_{pJ^k}^k\}$. Let $\boldsymbol{\lambda}_q^k = \{\lambda_{0,q}^k, \lambda_{1,q}^k, \ldots, \lambda_{j,q}^k, \ldots, \lambda_{J^k,q}^k, \lambda_{.,.,q}^k, \ldots, \lambda_{.,.,q}^k\}$, that is a cluster specific vector containing an intercept, all main effects and (a subset of) the two-way interaction effects. Let the number of elements of $\boldsymbol{\lambda}_q^k$ be denoted by L^k. Let $\boldsymbol{R}^k$ denotes a $2^{J^k} \times L^k$ design matrix. Then

$$P(\boldsymbol{x}_i^k \mid \tau_i = q) = \frac{exp^{\boldsymbol{R}_p^k \boldsymbol{\lambda}_q^k}}{\sum_{p'=1}^{2^{J^*}} exp^{\boldsymbol{R}_{p'}^k \boldsymbol{\lambda}_q^k}}, \tag{2.5}$$

where $\boldsymbol{R}_p^k$ denotes the row from $\boldsymbol{R}^k$ for which $\boldsymbol{x}_i^k = \boldsymbol{y}_p^k$. The interested reader is referred to, for example, Schafer (1997, p. 289-292) for a further and more general elaboration. We restrict ourselves to a simple elaboration with a set $\boldsymbol{X}^k$ with $J^k = 3$ and $\boldsymbol{\lambda}_q^k = \{\lambda_{0,q}^k, \lambda_{1,q}^k, \lambda_{2,q}^k, \lambda_{3,q}^k, \lambda_{1,2,q}^k, \lambda_{1,3,q}^k\}$, that is $L^k = 6$. Here

$$\boldsymbol{Y}^k = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

and

$$\boldsymbol{R}^k = \begin{pmatrix} 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

This example returns for further illustration in the next subsection.

The prior distribution in our model based cluster model is based on standard uninformative and mutually independent Dirichlet distributions for the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\omega}$. Stated otherwise, this prior has a density that is constant and independent of the values $\boldsymbol{\pi}$ and $\boldsymbol{\omega}$. As is elaborated in the next subsection, the log-linear parameters $\boldsymbol{\lambda}_q^k$ are sampled using the probabilities $P(\boldsymbol{y}_p^k \mid \boldsymbol{\tau} = q)$, for $p = 1, \dots, 2^{J^k}$, and the design matrix $\boldsymbol{R}^k$. Consequently a prior distribution has to be specified for these probabilities. A standard uninformative Dirichlet distribution is used (Schafer, 1997, p. 306) as the prior distribution for $P(\boldsymbol{y}_p^k \mid \boldsymbol{\tau} = q)$. Consequently, the prior distribution for the cluster model becomes:

$$h(\boldsymbol{\pi}, P(\boldsymbol{y}_p^1 \mid \boldsymbol{\tau} = 1), \dots, P(\boldsymbol{y}_p^K \mid \boldsymbol{\tau} = Q), \boldsymbol{\omega}) \propto \mathit{constant}. \tag{2.6}$$

The prior $h(\boldsymbol{\pi}, \boldsymbol{\lambda}(.)_1^1, \dots, \boldsymbol{\lambda}(.)_Q^K, \boldsymbol{\omega})$ follows from (2.6). Note that $\boldsymbol{\lambda}(.)_q^k$ denotes that the log-linear parameters for the k^{th} subset in cluster q are a function of the probabilities $P(\boldsymbol{y}_p^k \mid \boldsymbol{\tau} = q)$, for $p = 1, \dots, 2^{J^k}$.

The complete data posterior distribution of the cluster model is proportional to the product of the likelihood and the prior distribution:

$$Post(\boldsymbol{\pi}, \boldsymbol{\lambda}(\boldsymbol{.})_1^1, \ldots, \boldsymbol{\lambda}(\boldsymbol{.})_Q^K, \boldsymbol{\omega} \mid \boldsymbol{X}) \propto \tag{2.7}$$

$$L(\boldsymbol{X} \mid \boldsymbol{\pi}, \boldsymbol{\lambda}(\boldsymbol{.})_1^1, \ldots, \boldsymbol{\lambda}(\boldsymbol{.})_Q^K, \boldsymbol{\omega}) \times h(\boldsymbol{\pi}, \boldsymbol{\lambda}(\boldsymbol{.})_1^1, \ldots, \boldsymbol{\lambda}(\boldsymbol{.})_Q^K, \boldsymbol{\omega}).$$

Using the hierarchical clustering algorithm, that is described in the next two sections, it is easy to obtain a sample from the global mode of this posterior distribution.

2.3.3 The Gibbs Sampler

Markov Chain Monte Carlo (MCMC) is a class of methods that can be used to obtain a sample from posterior distributions (Gelman et al., 2000, p. 285-287; Schafer, 1997, p. 68-80; Zeger and Karim, 1991). In this chapter we use a particular MCMC method, that is Gibbs sampling. Gibbs sampling is a popular MCMC method and has been found useful in many multi-dimensional problems (Gelman et al., 2000, p. 287; Ter Braak et al., 2003). In Gibbs sampling the set of unknown parameters is split into a number of subsets. In each Gibbs iteration $z = 1, \ldots, Z$, each subset of parameters is sampled conditional on the most recently sampled values of all other subsets.

For the latent cluster model at hand we distinguish the following subsets of parameters:

- $\boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_Q\}$, that are the cluster specific probabilities for the J^0 locally independent items in each of the Q latent clusters.
- $\boldsymbol{\lambda}(\boldsymbol{.}) = \{\boldsymbol{\lambda}(\boldsymbol{.})_1^1, \ldots, \boldsymbol{\lambda}(\boldsymbol{.})_Q^K\}$, that are the cluster specific log-linear parameters for the K sets of locally dependent items in each of the Q latent clusters.
- $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_Q\}$, that are the cluster weights.
- $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N)$, that are the cluster memberships.
- $\{x_{ij} \mid m_{ij} = 1\}$, that are the missing values.

The last two subsets of parameters, the unobserved cluster memberships $\boldsymbol{\tau}$ and the missing values $\{x_{ij} \mid m_{ij} = 1\}$ are so called nuisance parameters. Sampling from a posterior containing nuisance parameters can be achieved using data augmentation (Hoijtink, 2000; Zeger and Karim, 1991). The structure of the resulting Gibbs sampling algorithm is described below, it consists of five steps preceded by an initialization. To initialize the Gibbs sampler a reasonable allocation of the respondents to the latent clusters is needed. How this is done is explained in Section 2.3.4. Furthermore, the elements of the set $\{x_{ij} \mid m_{ij} = 1\}$ are set to 1. After this initialization the iterations of the Gibbs sampler are started:

1. For $q = 1, \ldots, Q$ and $j = 1, ..., J^0$ sample cluster specific probability $\pi_{j|q}$ from $Post(\pi_{j|q} | \boldsymbol{x}_j, \boldsymbol{\tau})$. This is a $Dirichlet(\pi_{j|q} \mid N(\boldsymbol{x}_j = 0 \mid q) + 1, N(\boldsymbol{x}_j = 1 \mid q) + 1)$, where $N(\boldsymbol{x}_j = 0 \mid q)$ denotes the number of respondents who did not pick item j and are currently allocated to cluster q and $N(\boldsymbol{x}_j = 1 \mid q)$ denotes the number of respondents who did pick item j and are currently allocated to cluster q.

2. For $q = 1, \ldots, Q$ and $k = 1, \ldots, K$ sample cluster specific log-linear parameters $\boldsymbol{\lambda}(.)_q^k$ from $Post(\boldsymbol{\lambda}(.)_q^k \mid \boldsymbol{X}^k, \boldsymbol{\tau})$. This is achieved using Bayesian iterative proportional fitting (BIPF) (Gelman et al., 2000, p. 435-437; Schafer, 1997, p. 308-309). How BIPF works, is illustrated continuing the simple elaboration from the previous subsection.

   Let f_{abc}, for $a, b, c \in \{0, 1\}$, denotes the frequency with which each element of $\boldsymbol{Y}^k$ is observed in cluster q. Let θ_{abc} denotes the probability that a person in cluster q responds abc to the three items. To avoid heavy notation the subscript q and superscript k are implicit for f_{abc} and θ_{abc}. Let $\boldsymbol{\theta} = (\theta_{000}, \ldots, \theta_{111})$.

   The first time the Gibbs sampler enters Step 2 of the five step iterative procedure, $\theta_{abc} = \frac{1}{2^{J^k}}$, for $a, b, c \in \{0, 1\}$. In all other iterations Step 2 consists of two sub-steps:

   (a) Sample g_{ab+} from a standard Gamma distribution with shape parameters $f_{ab+} + 1$, for $a, b \in \{0, 1\}$, where $f_{ab+} = \sum_c f_{abc}$. Let $g_{+++} = \sum_{ab} g_{ab+}$, then
   $$\theta_{abc}^{new} = \theta_{abc}^{current} \left(\frac{g_{ab+}/g_{+++}}{\theta_{ab+}^{current}} \right) \text{ , for } a, b, c \in \{0, 1\}. \tag{2.8}$$

   (b) Sample g_{a+c} from a standard Gamma distribution with shape parameters $f_{a+c} + 1$, for $a, c \in \{0, 1\}$, where $f_{a+c} = \sum_b f_{abc}$. Let $g_{+++} = \sum_{ac} g_{a+c}$, then
   $$\theta_{abc}^{new} = \theta_{abc}^{current} \left(\frac{g_{a+c}/g_{+++}}{\theta_{a+c}^{current}} \right) \text{ , for } a, b, c \in \{0, 1\}. \tag{2.9}$$

   After execution of the two sub-steps of Step 2 the parameters of the log-linear model for subset k in cluster q are computed using

$$\boldsymbol{\lambda}_q^k = ((\boldsymbol{R}^k)^T \boldsymbol{R}^k)^{-1} (\boldsymbol{R}^k)^T log\boldsymbol{\theta}. \tag{2.10}$$

(Schafer, 1997, p. 299)

3. Sample the cluster weights $\boldsymbol{\omega}$ from $Post(\boldsymbol{\omega} \mid \boldsymbol{\tau})$. This is a $Dirichlet(\omega_q \mid N(\boldsymbol{\tau}=1)+1, \ldots, N(\boldsymbol{\tau}=q)+1, \ldots, N(\boldsymbol{\tau}=Q)+1)$ for $q = 1, \ldots, Q$, where $N(\boldsymbol{\tau}=q)$ denotes the number of respondents currently allocated to cluster q. See Narayanan (1990) for an overview of Dirichlet sampling methods.

4. For $i = 1, \ldots, N$ sample the respondents unobserved cluster memberships τ_i. This is a $Multinomial(\tau_i \mid P_{1|i}, \ldots, P_{q|i}, \ldots, P_{Q|i})$, where

$$P_{q|i} = \frac{P(\boldsymbol{x}_i \mid \tau_i = q)\omega_q}{\sum_{q'=1}^{Q} P(\boldsymbol{x}_i \mid \tau_i = q')\omega_{q'}}. \tag{2.11}$$

5. Sample each element x_{ij} from the set of missing values $\{x_{ij} \mid m_{ij} = 1\}$ sequentially. This is a *Bernouilli* distribution with a "success" probability that is calculated as follows:

   - if item j is a locally independent item, the "success" probability is

   $$P(x_{ij} = 1 \mid m_{ij} = 1, \boldsymbol{\pi}_q, \tau_i = q) = \pi_{j|q}, \tag{2.12}$$

   - if item j is a locally dependent item, the "success" probability is

   $$P(x^*_{ij} = 1 \mid m_{ij} = 1, \boldsymbol{\lambda}(.)_q, \tau_i = q) = \frac{exp^{\boldsymbol{R}_p \boldsymbol{\lambda}(.)_q}}{exp^{\boldsymbol{R}_p \boldsymbol{\lambda}(.)_q} + exp^{\boldsymbol{R}_t \boldsymbol{\lambda}(.)_q}}. \tag{2.13}$$

   where $\boldsymbol{R}_p$ is the row from $\boldsymbol{R}$ for which
   $y_{p1} = x_{i1}, \ldots, y_{pj} = 1, \ldots, y_{pJ^*} = x_{iJ^*}$ and $\boldsymbol{R}_t$ is the row from $\boldsymbol{R}$ for which $y_{t1} = x_{i1}, \ldots, y_{tj} = 0, \ldots, y_{tJ^*} = x_{iJ^*}$.

For each of the analysis to be executed in this chapter the number of Gibbs iterations Z is set to 1100. The number of Gibbs iterations is kept relatively small because for each cluster a separate Gibbs sampling algorithm is run. The first 100 iterations, $z = 1, \ldots, 100$, serve as burn-in iterations. These burn-in iterations are discarded to diminish the effect of the initial values. The last 1000 iterations, $z = 101, \ldots, 1100$, are the actual Gibbs iterations. For a number of the analysis executed convergence of the augmented Gibbs sampler was investigated (for an extensive discussion of convergence see Cowles and Carlin (1996)). It turned out that for the Gibbs sampling algorithm $Z = 1100$ with 100 burn-in iterations is sufficiently large to obtain convergence. In other words, the remaining sample of 1000 iterations constitutes a sample from the posterior (2.7).

2.3.4 How to Deal with Large Data Sets

Hoijtink and Notenboom (2004) present two conjectures with respect to the geometry of the posterior distribution of a standard cluster model. From these conjectures they derive a hierarchical cluster algorithm. There is no proof that the algorithm always renders the correct cluster structure, i.e. it is a heuristic algorithm. However, it is well motivated and in simulation studies it has been shown that this algorithm more often than not is able to reproduce the cluster structure in the population from which data were simulated. They also explain why for large data sets the algorithm is not sensitive for label switching (Stevens, 2000) and how their conjectures imply that the algorithm renders the global mode of the posterior distribution.

In this section their algorithm will shortly be presented and its workings will be illustrated for models with two-way interactions in a small example. Note that the algorithm of Hoijtink and Notenboom (2004) is related to and inspired by the algorithm presented by Richardson and Green (1997). The main difference is that Richardson and Green (1997) analyze the scores on one variable, while Hoijtink and Notenboom (2004) analyze large data sets containing many variables.

Let Q_{max} denotes the maximum number of clusters for (a subset of) the data. In the very first iteration of the hierarchical algorithm the whole sample of respondents is randomly split into $Q = 2$ clusters. Subsequently the Gibbs sampler described in the previous section is applied to allocate each person to one of the two clusters.

In all subsequent iterations of the hierarchical algorithm, the algorithm:

1. determines which cluster is the largest of the Q clusters at hand,

2. randomly splits the respondents from the largest cluster into two clusters,

3. applies the Gibbs sampling algorithm to these two clusters only, to determine for each person two which cluster he belongs,

4. applies the Gibbs sampling algorithm to all $Q + 1$ current clusters to allow between cluster migration.

Table 2.1: Items in the Simulation.

Sets of dependent items		Independent items
$k = 1$	var_1, var_2	var_3
$k = 2$	var_5, var_6	var_4
$k = 3$	var_9, var_{10}	var_7
$k = 4$	var_{16}, var_{17}	var_8
$k = 5$	var_{18}, var_{19}	var_{11}
$k = 6$	var_{20}, var_{21}	var_{12}
		var_{13}
		var_{14}
		var_{15}

If the resulting number of respondents in each of the $Q+1$ clusters is at least one, a new iteration is started. If at least one of the resulting $Q+1$ clusters is empty, the current iteration is repeated by splitting the next largest cluster. If there is no next largest cluster left, the hierarchical algorithm stops. The result of the hierarchical algorithm is a sample from the global mode of the posterior distribution for Q_{max} clusters. Note that Q_{max} does not have to be pre-specified. It is an outcome of the hierarchical algorithm and is an estimate of the number of clusters in the population from which the data are sampled.

To illustrate the hierarchical algorithm, a data set containing 400 respondents, 21 main effects and 6 interaction effects is simulated. The items assuming local within cluster dependencies are split into $K = 6$ sets of dependent items. See Table 2.1 for an overview of the dependent and independent items. In the simulated data set $Q_{max} = 3$. With $Q_{max} = 3$ the number of possible mixtures is five (one mixture of three clusters, three mixtures of two clusters in which two of the three clusters are combined and one mixture of one cluster). To illustrate that our model based cluster algorithm converges on the global mode of the posterior distribution, the cluster algorithm is run with three different initializations. In the first initialization Cluster 1 and 3 are combined, in the second initialization Cluster 2 and 3 are combined and in the third initialization Cluster 1 and 2 are combined. See Table 2.2 for the cluster specific parameters for $Q = 2$ for each of the three initializations.

Running the three initializations separately, each time the cluster algorithm stops after the third iteration. Stated otherwise, in all three runs $Q_{max} = 3$, which is the number of clusters simulated. See Table 2.3 for the cluster specific parameters for $Q_{max} = 3$ with the three initializations. As is shown in this table, the cluster specific parameters are more or less similar for the three initializations. This simi-

Table 2.2: Cluster Specific Parameters for the $Q = 2$ (or Local Mode) Solutions.

	Initialization 1		Initialization 2		Initialization 3	
Cluster q	1	2	1	2	1	2
Cluster weight ω_q	0.78	0.22	0.32	0.68	0.46	0.54
$\lambda(.)^1_{0,q}$	-1.73	-2.16	-2.36	-1.68	-2.22	-2.28
$\lambda(.)^1_{1,q}$	-0.08	-0.97	-1.05	0.06	0.66	-1.03
$\lambda(.)^1_{2,q}$	0.22	-0.88	-0.83	0.31	1.02	-0.87
$\lambda(.)^1_{1,2,q}$	0.85	0.01	0.30	0.71	0.31	0.15
$\lambda(.)^2_{0,q}$	-1.72	-2.46	-2.29	-2.26	-2.21	-1.66
$\lambda(.)^2_{1,q}$	-0.05	-0.94	0.75	-0.74	-0.71	0.16
$\lambda(.)^2_{2,q}$	-0.10	-1.43	1.00	-1.10	-0.95	0.04
$\lambda(.)^2_{1,2,q}$	0.84	-0.20	0.32	0.19	0.34	0.74
$\lambda(.)^3_{0,q}$	-2.26	-2.17	-2.60	-1.65	-2.14	-1.64
$\lambda(.)^3_{1,q}$	-0.90	0.86	-1.44	-0.09	-0.61	-0.24
$\lambda(.)^3_{2,q}$	-1.00	1.03	-1.17	-0.25	-0.98	0.01
$\lambda(.)^3_{1,2,q}$	0.13	-0.02	-0.24	0.68	0.29	0.68
$\lambda(.)^4_{0,q}$	-1.67	-2.35	-2.65	-1.70	-2.36	-2.52
$\lambda(.)^4_{1,q}$	0.04	-1.01	-1.33	0.12	0.99	-1.16
$\lambda(.)^4_{2,q}$	0.11	-0.94	-1.29	0.21	1.10	-1.14
$\lambda(.)^4_{1,2,q}$	0.77	0.20	-0.17	0.76	0.05	0.03
$\lambda(.)^5_{0,q}$	-1.70	-2.30	-2.37	-2.25	-2.21	-1.68
$\lambda(.)^5_{1,q}$	-0.20	-0.78	1.03	-0.95	-0.98	0.10
$\lambda(.)^5_{2,q}$	0.04	-1.06	1.09	-0.81	-0.67	0.07
$\lambda(.)^5_{1,2,q}$	0.80	0.24	0.04	0.28	0.35	0.78
$\lambda(.)^6_{0,q}$	-2.41	-2.21	-2.55	-1.64	-2.23	-1.62
$\lambda(.)^6_{1,q}$	-1.19	1.05	-1.35	-0.13	-0.96	-0.10
$\lambda(.)^6_{2,q}$	-1.12	0.75	-1.23	-0.22	-0.98	-0.10
$\lambda(.)^6_{1,2,q}$	-0.14	0.15	-0.29	0.66	0.03	0.68
$\pi_{1\|q}$	0.57	0.21	0.23	0.62	0.81	0.23
$\pi_{2\|q}$	0.55	0.22	0.19	0.61	0.80	0.20
$\pi_{3\|q}$	0.46	0.19	0.81	0.20	0.21	0.56
$\pi_{4\|q}$	0.44	0.23	0.79	0.21	0.20	0.56
$\pi_{5\|q}$	0.20	0.79	0.22	0.38	0.19	0.45
$\pi_{6\|q}$	0.19	0.82	0.24	0.38	0.17	0.47
$\pi_{7\|q}$	0.52	0.26	0.16	0.60	0.77	0.20
$\pi_{8\|q}$	0.45	0.28	0.84	0.21	0.18	0.61
$\pi_{9\|q}$	0.20	0.77	0.21	0.39	0.20	0.44

Table 2.3: Cluster Specific Parameters for the $Q_{max} = 3$ (or Global Mode) Solutions.

	Initialization 1			Initialization 2			Initialization 3		
Cluster q	1	2	3	1	2	3	1	2	3
Cluster weight ω_q	0.32	0.22	0.46	0.32	0.22	0.46	0.32	0.22	0.46
$\lambda(.)^1_{0,q}$	-2.36	-2.16	-2.20	-2.35	-2.16	-2.20	-2.35	-2.17	-2.20
$\lambda(.)^1_{1,q}$	-1.04	-0.96	0.65	-1.05	-0.97	0.64	-1.04	-0.97	0.64
$\lambda(.)^1_{2,q}$	-0.84	-0.88	1.00	-0.82	-0.88	1.00	-0.83	-0.89	1.00
$\lambda(.)^1_{1,2,q}$	0.29	0.01	0.32	0.29	0.01	0.33	0.29	0.00	0.32
$\lambda(.)^2_{0,q}$	-2.34	-2.47	-2.22	-2.34	-2.49	-2.21	-2.34	-2.48	-2.21
$\lambda(.)^2_{1,q}$	0.80	-0.94	-0.71	0.81	-0.95	-0.71	0.81	-0.95	-0.71
$\lambda(.)^2_{2,q}$	1.07	-1.45	-0.95	1.06	-1.46	-0.94	1.06	-1.45	-0.94
$\lambda(.)^2_{1,2,q}$	0.27	-0.20	0.34	0.27	-0.20	0.34	0.27	-0.20	0.35
$\lambda(.)^3_{0,q}$	-2.62	-2.19	-2.16	-2.61	-2.18	-2.16	-2.59	-2.18	-2.16
$\lambda(.)^3_{1,q}$	-1.43	0.86	-0.64	-1.42	0.84	-0.64	-1.40	0.85	-0.64
$\lambda(.)^3_{2,q}$	-1.18	1.04	-1.00	-1.18	1.03	-1.00	-1.16	1.04	-1.00
$\lambda(.)^3_{1,2,q}$	-0.22	-0.02	0.26	-0.22	0.00	0.27	-0.20	0.00	0.27
$\lambda(.)^4_{0,q}$	-2.62	-2.35	-2.26	-2.64	-2.36	-2.26	-2.66	-2.35	-2.26
$\lambda(.)^4_{1,q}$	-1.31	-1.01	0.90	-1.33	-1.02	0.90	-1.35	-1.00	0.90
$\lambda(.)^4_{2,q}$	-1.26	-0.92	0.98	-1.27	-0.95	0.98	-1.30	-0.92	0.99
$\lambda(.)^4_{1,2,q}$	-0.15	0.22	0.15	-0.16	0.20	0.15	-0.18	0.23	0.14
$\lambda(.)^5_{0,q}$	-2.38	-2.32	-2.21	-2.38	-2.33	-2.20	-2.38	-2.33	-2.22
$\lambda(.)^5_{1,q}$	1.00	-0.80	-0.99	1.01	-0.82	-0.99	1.01	-0.82	-1.00
$\lambda(.)^5_{2,q}$	1.10	-1.09	-0.66	1.10	-1.09	-0.65	1.10	-1.09	-0.66
$\lambda(.)^5_{1,2,q}$	0.05	0.21	0.34	0.05	0.20	0.34	0.06	0.20	0.34
$\lambda(.)^6_{0,q}$	-2.56	-2.25	-2.26	-2.56	-2.25	-2.27	-2.56	-2.24	-2.27
$\lambda(.)^6_{1,q}$	-1.37	1.08	-1.00	-1.37	1.08	-1.01	-1.36	1.07	-1.01
$\lambda(.)^6_{2,q}$	-1.22	0.79	-1.01	-1.21	0.80	-1.01	-1.21	0.81	-1.02
$\lambda(.)^6_{1,2,q}$	-0.27	0.11	-0.01	-0.28	0.11	-0.01	-0.27	0.10	-0.01
$\pi_{1\|q}$	0.23	0.21	0.81	0.23	0.21	0.81	0.23	0.21	0.81
$\pi_{2\|q}$	0.19	0.22	0.80	0.19	0.22	0.80	0.19	0.22	0.80
$\pi_{3\|q}$	0.82	0.19	0.21	0.82	0.19	0.21	0.82	0.19	0.21
$\pi_{4\|q}$	0.79	0.22	0.21	0.79	0.22	0.21	0.79	0.22	0.20
$\pi_{5\|q}$	0.21	0.79	0.19	0.22	0.79	0.19	0.21	0.79	0.19
$\pi_{6\|q}$	0.23	0.82	0.17	0.23	0.82	0.17	0.23	0.82	0.17
$\pi_{7\|q}$	0.16	0.26	0.77	0.16	0.26	0.77	0.16	0.26	0.77
$\pi_{8\|q}$	0.84	0.28	0.18	0.84	0.28	0.18	0.84	0.28	0.18
$\pi_{9\|q}$	0.21	0.77	0.20	0.21	0.77	0.20	0.21	0.77	0.20

larity in cluster specific parameters for the three initializations is an indication that the cluster algorithm has converged on the global mode of the posterior distribution, or, in other words, $Q_{max} = 3$. That the $Q = 2$ solutions are local mode solutions with three possible mixtures of two cluster in which two of the three clusters from the $Q_{max} = 3$ solution are combined, can be seen in Table 2.3. Cluster 1 in the $Q = 2$ solution in Initialization 1 (Table 2.2) is a combination of Cluster 1 and 3 in the $Q_{max} = 3$ solution in Initialization 1 (Table 2.3). This shows that the conjectures are practically useful and that the hierarchical algorithm converges to the global mode of the posterior distribution.

2.4 Determining the Number of Latent Clusters

2.4.1 Introduction

The question 'how many clusters?' is one of the main research topics in model based clustering. According to the two conjectures the number of clusters in the data set at hand is equal to Q_{max}. However, this is not necessarily the number of clusters in the population from which the data set is coming. No clear procedures and/or guidelines to determine the number of clusters for large data sets exist. For small data sets usually two approaches are combined to estimate the number of clusters in the population: information criteria (Lin and Dayton, 1997; Vermunt and Magidson, 2000, p. 61) and likelihood ratio tests (Everitt, 1988; Vermunt and Magidson, 2000, p. 61). The next subsection describes the pseudo-likelihood ratio test which is used in this chapter and tailored for use with large data sets.

2.4.2 Pseudo-Likelihood Ratio Test

In this chapter a likelihood ratio test is used to determine the number of latent clusters for the data set at hand. Likelihood ratio tests are absolute fit measures. These measures rely upon a comparison between what is observed in the data set and what is expected given the cluster model parameters. Latent cluster models that lead to expectations that are too far from what is observed, are deemed unacceptable or implausible, whereas models that yield expectations that are similar to what has actually been observed are believed to be more plausible or acceptable (Hagenaars and McCutcheon, 2002, p. 66-69). However, likelihood ratio tests are useless if the number of items in the data set is large. Had the data been complete, the number of possible response vectors would have been 2^{149}. It is clear that even with a sample size of $N = 2294$ this constitutes a very sparse contingency table. In the complete data situation the null-distribution of the likelihood ratio test is unclear due to this sparseness (see, for example, Agresti, 2002, p. 246-247). Our situation is complicated further by the presence of data that are missing.

A test statistic that can both handle large data sets and deals with missing values is the pseudo-likelihood ratio (PLR) test (Hoijtink, 1998). This test is a discrepancy measure (Gelman et al., 1996; Meng, 1994), that is a test statistic that is a function of both the data and the unknown model parameters.

$$D(\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{\lambda(.)}, \boldsymbol{\omega}) =$$

$$-2 \sum_{j \neq j'} \sum_{u=0}^{1} \sum_{v=0}^{1} N(x_j = u, x_{j'} = v) \log\left(\frac{E(x_j = u, x_{j'} = v \mid \boldsymbol{\pi}, \boldsymbol{\lambda(.)}, \boldsymbol{\omega})}{N(x_j = u, x_{j'} = v)}\right), \tag{2.14}$$

where

$$E(x_j = u, x_{j'} = v \mid \boldsymbol{\pi}, \boldsymbol{\lambda(.)}, \boldsymbol{\omega}) = \sum_{q=1}^{Q} P(x_j = u, x_{j'} = v \mid q)\left(\sum_{i \mid m_{ij}=0, m_{ij'}=0} P_{q|i}\right). \tag{2.15}$$

In (2.14) $N(x_j = u, x_{j'} = v)$ denotes the observed number of respondents answering u on item j and v on item j', where the item responses $u, v \in \{0, 1\}$. And, (2.15) denotes the expected number of respondents answering u on item j and v on item j'. The probability $P(x_j = u, x_{j'} = v \mid q)$ in (2.15) is calculated as follows:

- if item j and j' are locally independent items,

$$P(x_j = u, x_{j'} = v \mid q) = P(x_j = u \mid q) P(x_{j'} = v \mid q), \tag{2.16}$$

with

$$P(x_j = u \mid q) = \pi_{j|q}^{u} (1 - \pi_{j|q})^{1-u}, \tag{2.17}$$

and

$$P(x_{j'} = v \mid q) = \pi_{j'|q}^{v} (1 - \pi_{j'|q})^{1-v}. \tag{2.18}$$

- if item j is a locally independent items and item j' is a locally dependent item in the k^{th} set of dependent items,

$$P(x_j = u, x_{j'} = v \mid q) = P(x_j = u \mid q) P(x_{j'} = v \mid q), \tag{2.19}$$

with

$$P(x_j = u \mid q) = \pi_{j|q}^{u} (1 - \pi_{j|q})^{1-u}, \tag{2.20}$$

and

$$P(x_{j'} = v \mid q) = \frac{\sum_{p \mid y_{pj'}^{k} = v} exp^{\boldsymbol{R}_{p}^{k} \boldsymbol{\lambda(.)}_{q}^{k}}}{\sum_{p'=1}^{2^{J^k}} exp^{\boldsymbol{R}_{p'}^{k} \boldsymbol{\lambda(.)}_{q}^{k}}}. \tag{2.21}$$

- if item j and j' are locally dependent items and both item j and j' are in the k^{th} set of dependent items,

$$P(x_j = u, x_{j'} = v \mid q) = \frac{\sum_{p|y^k_{pj}=u, y^k_{pj'}=v} exp^{\boldsymbol{R}^k_p \boldsymbol{\lambda}(.)^k_q}}{\sum_{p'=1}^{2^{J^k}} exp^{\boldsymbol{R}^k_{p'} \boldsymbol{\lambda}(.)^k_q}}. \tag{2.22}$$

- if item j and j' are locally dependent items and item j is in the k^{th} set of dependent items and item j' is in the $k^{'th}$ set of dependent items,

$$P(x_j = u, x_{j'} = v \mid q) = P(x_j = u \mid q) P(x_{j'} = v \mid q), \tag{2.23}$$

  with

$$P(x_j = u \mid q) = \frac{\sum_{p|y^k_{pj}=u} exp^{\boldsymbol{R}^k_p \boldsymbol{\lambda}(.)^k_q}}{\sum_{p'=1}^{2^{J^k}} exp^{\boldsymbol{R}^k_{p'} \boldsymbol{\lambda}(.)^k_q}}, \tag{2.24}$$

  and

$$P(x_{j'} = v \mid q) = \frac{\sum_{p|y^{k'}_{pj'}=v} exp^{\boldsymbol{R}^k_p \boldsymbol{\lambda}(.)^{k'}_q}}{\sum_{p'=1}^{2^{J^{k'}}} exp^{\boldsymbol{R}^{k'}_{p'} \boldsymbol{\lambda}(.)^{k'}_q}}. \tag{2.25}$$

The term $\sum_{i|m_{ij}=0, m_{ij'}=0} P_{q|i}$ in (2.15) denotes the number of respondents assigned to latent cluster q, who do not have missing values on both items j and j'.

As is shown in (2.14), the pseudo-likelihood ratio test focuses on two-dimensional summaries of expected and observed frequencies of the J-dimensional contingency table. This implies that the test only evaluates whether the main effects and two-way interactions are adequately predicted. According to our model choice it is safe to state that a good cluster model should be able to predict the main effects and two-way interaction effects. Hoijtink (1998, 2001) did some simulations indicating that for sparse contingency tables the pseudo-likelihood ratio test has a better performance than the likelihood ratio test.

2.4.3 How to Deal with Missing Values

The pseudo-likelihood ratio test presented in the previous section can be used to evaluate the fit of a mixture of log-linear models containing only main effects and two-way interactions. Data expunction (Hoijtink and Notenboom, 2004) will be used to compute a p-value for this test. Formal hypothesis testing in the presence of missing values is usually based on multiple imputation (Little and Rubin, 2002; Rubin, 1987; Schafer, 1997; Schafer and Graham, 2002). Multiple imputation cannot be used to evaluate goodness of fit tests that address "fixed" properties of a model like normality, linearity and homoscedasticity. Therefore it is not possible to determine with multiple imputation how good the fit of a model with Q latent clusters is. With data expunction it is possible to determine the goodness of fit.

The difference between the method of multiple imputation and the method of data expunction is that in the first method the missing values in the observed data set $\boldsymbol{X}$ are imputed and in the latter method data are "expuncted" from data sets replicated from the null population, denoted by $\boldsymbol{X}^{rep}$. In both methods statistical inference can only be made if the missing value mechanism is known. For the data set at hand a missing value mechanism is proposed in the following subsection.

To evaluate the pseudo-likelihood ratio discrepancy measure in (2.14) posterior predictive p-values (Gelman et al., 2000, p. 167-173; Meng, 1994) are used.

$$P(D(\boldsymbol{X}^{rep}, \boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{\lambda}(.), \boldsymbol{\omega}) \geq D(\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{\lambda}(.), \boldsymbol{\omega}) \mid \boldsymbol{X}, \boldsymbol{M}). \tag{2.26}$$

This can only be done if the missing value mechanism is explicitly accounted for. The probability (2.26) is evaluated with respect to the distribution of the five random variables $\boldsymbol{X}^{rep}$, $\boldsymbol{\pi}$, $\boldsymbol{\lambda}(.)$, $\boldsymbol{\omega}$ and $\boldsymbol{\Phi}$:

$$\begin{gathered} g(\boldsymbol{X}^{rep}, \boldsymbol{\pi}, \boldsymbol{\lambda}(.), \boldsymbol{\omega}, \boldsymbol{\Phi} \mid \boldsymbol{X}, \boldsymbol{M}) = \\ g(\boldsymbol{X}^{rep} \mid \boldsymbol{\pi}, \boldsymbol{\lambda}(.), \boldsymbol{\omega}, \boldsymbol{\Phi}) \mathrm{Post}(\boldsymbol{\pi}, \boldsymbol{\lambda}(.), \boldsymbol{\omega} \mid \boldsymbol{X}, \boldsymbol{M}) \mathrm{Post}(\boldsymbol{\Phi} \mid \boldsymbol{X}, \boldsymbol{M}). \end{gathered} \tag{2.27}$$

In (2.27) the parameter vector $\boldsymbol{\Phi}$ represents the missing value mechanism. It is described in the next subsection. A four-step simulation is used to actually compute (2.26):

1. For each Gibbs iteration $z = 1, \ldots, Z$ sample $\boldsymbol{\pi}$, $\boldsymbol{\lambda}$, $\boldsymbol{\omega}$ and $\boldsymbol{\tau}$ from (2.7) using the Gibbs sampler as described in Section 2.3.3.

2. For each Gibbs iteration $z = 1, \ldots, Z$ sample $\boldsymbol{X}^{rep}$ from (2.1). This is done in two sub-steps: (i) sample the complete $\boldsymbol{X}^{rep}$ using the parameters from the previous step and (ii) apply the missing value mechanism as described in the next subsection.

3. For each Gibbs iteration $z = 1, \ldots, Z$ compute $D(\boldsymbol{X}^{rep}, \boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{\lambda}(.), \boldsymbol{\omega})$ and $D(\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{\lambda}(.), \boldsymbol{\omega})$.

4. Compute the proportion $D(\boldsymbol{X}^{rep}, \boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{\lambda}(.), \boldsymbol{\omega}) > D(\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{\lambda}(.), \boldsymbol{\omega})$.

Table 2.4: Missing Question Pattern for Data Set at hand.

s	missing pattern	ϕ_s		s	missing pattern	ϕ_s
1	00000	0.9058		17	10000	0.0100
2	00001	0.0113		18	10001	0.0022
3	00010	0.0126		19	10010	0.0026
4	00011	0.0017		20	10011	0.0022
5	00100	0.0065		21	10100	0.0035
6	00101	0.0026		22	10101	0.0009
7	00110	0.0017		23	10110	0.0017
8	00111	0.0009		24	10111	0.0013
9	01000	0.0092		25	11000	0.0026
10	01001	0.0044		26	11001	0.0013
11	01010	0.0031		27	11010	0.0013
12	01011	0.0013		28	11011	0.0004
13	01100	0.0035		29	11100	0.0004
14	01101	0.0009		30	11101	0.0004
15	01110	0.0009		31	11110	0.0013
16	01111	0.0013		32	11111	0.0000

Missing value mechanism

As described in Section 2.2 the data set is a collection of five questions from the so-called BSR questionnaire (see Appendix I for this questionnaire). Each question contains multiple psychographic items. The respondent is asked to characterize a person, resembling himself, who looks or feels the same as the respondent towards a particular domain. For each of the five questions the respondent is asked to pick the items which describe the person he has in mind the best. Because each question contains a broad range of items it is unlikely that a respondent can not pick an item from the item list. If it occurs that a respondent did not pick an item from the item list, it is assumed that the respondent skipped the whole question accidently. All items from this particular question are then considered to be missing values. From this description it is clear that a respondent can have zero up to all five questions missing. As such there are $s = 1, \ldots, 2^5$ possible missing question patterns (see Table 2.4).

In Table 2.4 $s = 1$ is 00000, indicating that there are no questions missing, $s = 2$ is 00001, indicating that only the last question is missing, and so on. Let $\mathbf{\Phi} = \{\phi_1, \ldots, \phi_s, \ldots, \phi_{32}\}$ denotes the parameters of the missing value mechanism.

That is each missing question pattern s has a probability ϕ_s to occur. This mechanism is missing completely at random because whether or not a response is missing is independent of other observed or unobserved variables. See Table 2.4 for the probabilities ϕ_s for the data set at hand. The missing value mechanism $\mathbf{\Phi}$ is formulated as follows. Let N^s denotes the observed number of respondents that has missing pattern $s = 1, \ldots, 32$. For each respondent the probabilities $\mathbf{\Phi}$ are sampled from $\text{Post}(\mathbf{\Phi} \mid \boldsymbol{X}, \boldsymbol{M})$. This is a $Dirichlet(\phi_s \mid N^1 + 1, \ldots, N^{32} + 1)$, for $s = 1, \ldots, 32$. The missing question pattern is sampled from a $Multinomial$ distribution with probabilities ϕ_s, for $s = 1, \ldots, 32$. Applying the result to the replicated data set $\boldsymbol{X}^{rep}$ for each respondent is called "data expunction".

2.5 Application

As explained in Section 2.2, the data set at hand contains $N = 2294$ respondents who answered to $J = 149$ psychographic items. See column 'Item' and 'Label' in Table 2.5 for an overview of the items in the data set and there corresponding labels. Before this large data set is put into the clustering algorithm, a model must be specified, or, in other words, it has to be decided which two-way interactions should be included in the model.

Based on prior knowledge about the underlying theoretical framework of BSR, the 90 two-way interactions in Table 2.6 are considered to be important to include in the model. All other items are considered to be locally independent items within the latent clusters. Column 'Type' in Table 2.5 shows whether an item is locally independent ('indep.') or locally dependent ('dep.'). Running the clustering algorithm with the specified model renders five clusters, or, in other words $Q_{max} = 5$. It turns out that with the data set at hand and the specified model it is not possible to create six clusters such that the expected number of respondents in each latent cluster is larger than one. The row 'ω_q' in Table 2.7 displays the expected a posteriori cluster weights for the $Q_{max} = 5$ solutions. From the row 'ω_q' it can be seen that Cluster 1 ($\omega_1 = 0.105$), 2 ($\omega_2 = 0.368$), 3 ($\omega_3 = 0.361$) and 4 ($\omega_4 = 0.164$) has relatively large cluster weights and therefore supposed to be substantial. Cluster 5 ($\omega_5 = 0.002$) has a relatively small cluster weight, representing only five respondents from the data set and is supposed to be not substantial. Furthermore, Table 2.7 shows for each item the item probability per cluster. These item probabilities, $P(x_{ij} = 1 \mid \tau_i = q)$, for $j = 1, \ldots, 149$, can be calculated as follows:

- if x_{ij} is a locally independent item,

$$P(x_{ij} = 1 \mid \tau_i = q) = P(x_{ij} = 1 \mid \boldsymbol{x}_i, \boldsymbol{\pi}_q, \tau_i = q) = \pi_{j|q}. \qquad (2.28)$$

Table 2.5: Items in the Data Set.

Type	Item	Label	Type	Item	Label	Type	Item	Label
indep.	x_{i1}	A little bit shy	dep.	x_{i56}	Member of the board	indep.	x_{i97}	Swimming
indep.	x_{i2}	Easygoing	dep.	x_{i57}	Financial planner	indep.	x_{i98}	Dine out together
indep.	x_{i3}	A little bit impatient	indep.	x_{i58}	No occupation	indep.	x_{i99}	Camping
indep.	x_{i4}	Honest	dep.	x_{i59}	Nurse	dep.	x_{i100}	Active sports
indep.	x_{i5}	Assertive	dep.	x_{i60}	Secretary	indep.	x_{i101}	Top-notch achievement
indep.	x_{i6}	Critical	indep.	x_{i61}	Vets assistant	indep.	x_{i102}	Going to a discotheque
indep.	x_{i7}	Interested in others	indep.	x_{i62}	Part time housewife	indep.	x_{i103}	Working out
indep.	x_{i8}	Gentle	indep.	x_{i63}	Shop assistant	indep.	x_{i104}	Gardening
indep.	x_{i9}	Jovial	dep.	x_{i64}	Receptionist	indep.	x_{i105}	A day out
indep.	x_{i10}	Neat	indep.	x_{i65}	Male nurse	indep.	x_{i106}	Golf
indep.	x_{i11}	Ordinary	dep.	x_{i66}	Sports teacher	indep.	x_{i107}	Surfing the Internet
indep.	x_{i12}	Capable	dep.	x_{i67}	Designer	indep.	x_{i108}	Team sports
indep.	x_{i13}	Spontaneous	indep.	x_{i68}	Volunteer	indep.	x_{i109}	Relaxing at home
indep.	x_{i14}	Strong character	dep.	x_{i69}	Director	indep.	x_{i110}	Shopping
dep.	x_{i15}	Adventurous	dep.	x_{i70}	Shopkeeper	indep.	x_{i111}	Going out together
indep.	x_{i16}	Sympathetic	dep.	x_{i71}	Project manager	indep.	x_{i112}	Do odd jobs around the house
indep.	x_{i17}	Energetic	dep.	x_{i72}	Commercial assistant	indep.	x_{i113}	Visiting a pub
indep.	x_{i18}	Self-confident	indep.	x_{i73}	Student	indep.	x_{i114}	Investing in stocks
dep.	x_{i19}	Leader	indep.	x_{i74}	Journalist	indep.	x_{i115}	Playing chess
indep.	x_{i20}	Classy	dep.	x_{i75}	Stylist	indep.	x_{i116}	Visiting friends and relatives
indep.	x_{i21}	Serious	indep.	x_{i76}	Truck driver	indep.	x_{i117}	Make dreams come through!
indep.	x_{i22}	A little impudent	dep.	x_{i77}	Business-man/-woman	indep.	x_{i118}	Watching tv
indep.	x_{i23}	Commercial	indep.	x_{i78}	Full time housewife	indep.	x_{i119}	Religious matters
indep.	x_{i24}	Cozy	indep.	x_{i79}	Freelancer	indep.	x_{i120}	Cars/ motor bikes
indep.	x_{i25}	Self-assured	indep.	x_{i80}	e-business	indep.	x_{i121}	Classy parties
indep.	x_{i26}	Deliberate	indep.	x_{i81}	Beauty specialist	indep.	x_{i122}	Build a successful career
dep.	x_{i27}	Passionate	indep.	x_{i82}	Anchor man	indep.	x_{i123}	A sociable evening with friends
indep.	x_{i28}	Serene	indep.	x_{i83}	Unemployed	indep.	x_{i124}	Snowboarding
indep.	x_{i29}	Intelligent	indep.	x_{i84}	Public servant	indep.	x_{i125}	Reading magazines
indep.	x_{i30}	Opinionated	dep.	x_{i85}	Social worker	indep.	x_{i126}	Astrology
indep.	x_{i31}	Helpful	indep.	x_{i86}	House-husband	dep.	x_{i127}	Adventurous holidays
indep.	x_{i32}	Down-to-earth	indep.	x_{i87}	Activity guide	indep.	x_{i128}	Self-belief
indep.	x_{i33}	Enthusiastic	indep.	x_{i88}	Temporary employee	indep.	x_{i129}	Self-advancement, growth
indep.	x_{i34}	Balanced	dep.	x_{i89}	Account manager	indep.	x_{i130}	Enthusiasm
indep.	x_{i35}	Cheerful	indep.	x_{i90}	Photographer	indep.	x_{i131}	Enjoying life
dep.	x_{i36}	Bachelor	dep.	x_{i91}	Manager	indep.	x_{i132}	Social harmony
dep.	x_{i37}	Happy family	indep.	x_{i92}	Programmer	indep.	x_{i133}	Friendship
indep.	x_{i38}	Not suited for family life	dep.	x_{i93}	Artist	indep.	x_{i134}	Social alliance
dep.	x_{i39}	Sportive family	dep.	x_{i94}	Welfare worker	dep.	x_{i135}	Passion
indep.	x_{i40}	Stable household	indep.	x_{i95}	Scientist	indep.	x_{i136}	Solidarity
indep.	x_{i41}	Isolated family	dep.	x_{i96}	Entrepreneur	indep.	x_{i137}	Intimacy
indep.	x_{i42}	Striving for a family				indep.	x_{i138}	Privacy, tranquility
indep.	x_{i43}	Harmonious familiy				indep.	x_{i139}	Respect
indep.	x_{i44}	Rigid family				indep.	x_{i140}	Security
indep.	x_{i45}	Aristocratic household				indep.	x_{i141}	Anonymity
dep.	x_{i46}	Single				indep.	x_{i142}	Rationality
dep.	x_{i47}	Busy, dynamical family				indep.	x_{i143}	Status
dep.	x_{i48}	Artistic household				indep.	x_{i144}	Recognition of performances
indep.	x_{i49}	A family everyone goes own way				indep.	x_{i145}	Heroism, glory
indep.	x_{i50}	Cozy old-fashioned family				dep.	x_{i146}	Being unique, different
indep.	x_{i51}	Warm family				indep.	x_{i147}	Independency
indep.	x_{i52}	Broad-minded family				indep.	x_{i148}	Success in life
indep.	x_{i53}	Ideal family				indep.	x_{i149}	Challenge
dep.	x_{i54}	Quiet family						
indep.	x_{i55}	Perfect family						

Table 2.6: Two-way Interaction Effects Included in the Model.

1	Interested in others (x_{i7})-Welfare worker (x_{i94})	46	Beauty specialist (x_{i81})-Shopping (x_{i110})
2	Financial planner (x_{i57})-Account manager (x_{i89})	47	Sportive family (x_{i39})-Team sports (x_{i108})
3	e-business (x_{i80})-Surfing the Internet (x_{i107})	48	Passionate (x_{i27})-Passion (x_{i135})
4	Cozy (x_{i24})-Self-assured (x_{i25})	49	Helpful (x_{i31})-Volunteer (x_{i68})
5	Truck driver (x_{i76})-Cars/ motor bikes (x_{i120})	50	Spontaneous (x_{i13})-Serious (x_{i21})
6	Top-notch achievement (x_{i101})-Heroism, glory (x_{i145})	51	Intelligent (x_{i29})-Self-advancement, growth (x_{i129})
7	Nurse (x_{i59})-A day out (x_{i105})	52	Member of the board (x_{i56})-Director (x_{i69})
8	Activity guide (x_{i87})-Welfare worker (x_{i94})	53	Bachelor (x_{i36})-Not suited for family life (x_{i38})
9	Manager (x_{i91})-Rationality (x_{i142})	54	Busy, dynamical family (x_{i47})-Quiet family (x_{i54})
10	Sports teacher (x_{i66})-Active sports (x_{i100})	55	Programmer (x_{i92})-Surfing the Internet (x_{i107})
11	Commercial assistant (x_{i72})-Account manager (x_{i89})	56	Enthusiastic (x_{i33})-Enthusiasm (x_{i130})
12	Build a successful career (x_{i122})-Success in life (x_{i148})	57	Social worker (x_{i85})-Welfare worker (x_{i94})
13	Relaxing at home (x_{i109})-Privacy, tranquility (x_{i138})	58	Stylist (x_{i75})-Beauty specialist (x_{i81})
14	Stylist (x_{i75})-Shopping (x_{i110})	59	Bachelor (x_{i36})-Single (x_{i46})
15	Sportive family (x_{i39})-Sports teacher (x_{i66})	60	Build a successful career (x_{i122})-Heroism, glory (x_{i145})
16	Nurse (x_{i59})-Welfare worker (x_{i94})	61	Artistic household (x_{i48})-Artist (x_{i93})
17	Happy family (x_{i37})-Striving for a family (x_{i42})	62	Nurse (x_{i59})-Vets assistant (x_{i61})
18	Capable (x_{i12})-Manager (x_{i91})	63	Ordinary (x_{i11})-Quiet family (x_{i54})
19	Bachelor (x_{i36})-Happy family (x_{i37})	64	Security (x_{i140})-Challenge (x_{i149})
20	Going to a discotheque (x_{i102})-Visiting a pub (x_{i113})	65	Leader (x_{i19})-Manager (x_{i91})
21	Classy (x_{i20})-Golf (x_{i106})	66	Bachelor (x_{i36})-Striving for a family (x_{i42})
22	Project manager (x_{i71})-Manager (x_{i91})	67	Happy family (x_{i37})-A family everyone goes own way (x_{i49})
23	Warm family (x_{i51})-Broad-minded family (x_{i52})	68	Sports teacher (x_{i66})-Team sports (x_{i108})
24	Nurse (x_{i59})-Social worker (x_{i85})	69	Shopkeeper (x_{i70})-Entrepreneur (x_{i96})
25	Artist (x_{i93})-Make dreams come through! (x_{i117})	70	Cozy old-fashioned family (x_{i50})-Full time housewife (x_{i78})
26	Secretary (x_{i60})-Receptionist (x_{i64})	71	Harmonious familiy (x_{i43})-Busy, dynamical family (x_{i47})
27	Financial planner (x_{i57})-Investing in stocks (x_{i114})	72	Social worker (x_{i85})-Activity guide (x_{i87})
28	Commercial (x_{i23})-Financial planner (x_{i57})	73	Spontaneous (x_{i13})-Cozy (x_{i24})
29	Full time housewife (x_{i78})-Manager (x_{i91})	74	Artistic household (x_{i48})-Being unique, different (x_{i146})
30	Project manager (x_{i71})-Full time housewife (x_{i78})	75	Nurse (x_{i59})-Freelancer (x_{i79})
31	Classy (x_{i20})-Aristocratic household (x_{i45})	76	Not suited for family life (x_{i38})-Single (x_{i46})
32	Leader (x_{i19})-Project manager (x_{i71})	77	Part time housewife (x_{i62})-Receptionist (x_{i64})
33	Designer (x_{i67})-Stylist (x_{i75})	78	No occupation (x_{i58})-Unemployed (x_{i83})
34	Going to a discotheque (x_{i102})-Passion (x_{i135})	79	Self-advancement, growth (x_{i129})-Friendship (x_{i133})
35	Ordinary (x_{i11})-Strong character (x_{i14})	80	Happy family (x_{i37})-Single (x_{i46})
36	Volunteer (x_{i68})-Social worker (x_{i85})	81	Adventurous (x_{i15})-Adventurous holidays (x_{i127})
37	Helpful (x_{i31})-Nurse (x_{i59})	82	Photographer (x_{i90})-Artist (x_{i93})
38	Receptionist (x_{i64})-Security (x_{i140})	83	Artist (x_{i93})-Being unique, different (x_{i146})
39	Stable household (x_{i40})-Warm family (x_{i51})	84	Intelligent (x_{i29})-Scientist (x_{i95})
40	Spontaneous (x_{i13})-Deliberate (x_{i26})	85	Business-man/-woman (x_{i77})-Entrepreneur (x_{i96})
41	Down-to-earth (x_{i32})-Enthusiastic (x_{i33})	86	Top-notch achievement (x_{i101})-Build a successful career (x_{i122})
42	Cozy (x_{i24})-Deliberate (x_{i26})	87	Volunteer (x_{i68})-Activity guide (x_{i87})
43	Honest (x_{i4})-Opinionated (x_{i30})	88	Shopkeeper (x_{i70})-Business-man/-woman (x_{i77})
44	e-business (x_{i80})-Programmer (x_{i92})	89	Shop assistant (x_{i63})-Project manager (x_{i71})
45	Financial planner (x_{i57})-Commercial assistant (x_{i72})	90	Sportive family (x_{i39})-Active sports (x_{i100})

- if item x_{ij} is a locally dependent item and item x_{ij} is in the k^{th} set of dependent items,

$$P(x_{ij}^k = 1 \mid \tau_i = q) = P(x_{ij}^k = 1 \mid \boldsymbol{x}_i^k, \boldsymbol{\lambda}(.)_q^k, \tau_i = q) = \frac{\sum_{p|y_{pj}^k=1} exp^{\boldsymbol{R}_p^k \boldsymbol{\lambda}(.)_q^k}}{\sum_{p'=1}^{2^{J^k}} exp^{\boldsymbol{R}_{p'}^k \boldsymbol{\lambda}(.)_q^k}}. \tag{2.29}$$

These probabilities are used in the cluster descriptions later in this section. The results of applying our clustering algorithm on the data set at hand are described using two of the six criteria from Wedel and Kamakura (2000, p. 4-5). These six criteria are frequently used to determine the effectiveness and profitability of market segmentation.

1. ***Substantiality***. This criterion is satisfied if the clusters represent a large enough portion of the market to ensure the profitability of targeted marketing programs. Substantiality is closely connected to the marketing goals (targeting micro markets vs. mass customization) and costs (with micro markets you have to make more marketing strategies). Although studying this criterion is not the purpose of this chapter, the row with cluster weights in Table 2.7 shows that there is a cluster with a very small cluster weight, that is Cluster 5 ($\omega_5 = 0.002$), representing only five respondents from the data set. Due to this small cluster weight this cluster is considered to be an outlier and not profitable. Therefore we focus on the four remaining clusters, which are considered to be profitable according to their cluster weights, that is Clusters 1 ($\omega_1 = 0.105$), 2 ($\omega_2 = 0.368$), 3 ($\omega_3 = 0.361$) and 4 ($\omega_4 = 0.164$).

2. ***Identifiability***. This criterion is the extent to which marketeers can recognize distinct groups of respondents. Using the item probabilities from Table 2.7 each of the four remaining latent clusters can be described in terms of probabilities. As described in Section 2.2, the idea behind the BSR framework is that there are four main motivational clusters, which has been found useful in marketing (Brethouwer et al., 1995, p. 50). All other clusters are considered to be combinations in terms of description of these four main clusters. Looking at our results, the four remaining clusters can be identified as these four main motivational clusters. For example, Cluster 3 corresponds with the cluster in the lower left quadrant in the BSR strategic map (see Figure 2.1). This cluster is described with the word 'Manifestation'. Persons from this cluster are career oriented and aspire a certain (high) status in life. Looking at Table 2.7, it can be seen that, for example, the items 'Intelligent'

Table 2.7: Cluster Specific Item Probabilities for the $Q_{max} = 5$ (or Global Mode) Solution. $P1 = P(x_{i1} = 1 \mid \tau_i = q)$,..., $P149 = P(x_{i149} = 1 \mid \tau_i = q)$

q	1	2	3	4	5	q	1	2	3	4	5	q	1	2	3	4	5
ω_q	0.105	0.368	0.361	0.164	0.002	ω_q	0.105	0.368	0.361	0.164	0.002	ω_q	0.105	0.368	0.361	0.164	0.002
$P1$	0.09	0.25	0.09	0.11	0.00	$P51$	0.35	0.33	0.35	0.47	0.00	$P101$	0.02	0.00	0.04	0.01	0.00
$P2$	0.13	0.10	0.15	0.23	0.00	$P52$	0.53	0.21	0.32	0.23	0.00	$P102$	0.02	0.01	0.05	0.10	0.00
$P3$	0.17	0.21	0.30	0.19	1.00	$P53$	0.02	0.01	0.04	0.06	1.00	$P103$	0.01	0.02	0.08	0.04	0.00
$P4$	0.47	0.65	0.50	0.61	1.00	$P54$	0.10	0.37	0.16	0.09	0.00	$P104$	0.48	0.41	0.22	0.22	0.00
$P5$	0.21	0.04	0.13	0.11	1.00	$P55$	0.00	0.01	0.03	0.03	0.00	$P105$	0.26	0.46	0.25	0.48	0.00
$P6$	0.45	0.19	0.37	0.12	0.00	$P56$	0.08	0.03	0.16	0.00	0.00	$P106$	0.01	0.01	0.07	0.01	0.00
$P7$	0.65	0.47	0.30	0.57	0.00	$P57$	0.05	0.08	0.26	0.06	1.00	$P107$	0.09	0.14	0.21	0.13	1.00
$P8$	0.10	0.21	0.11	0.21	0.00	$P58$	0.11	0.15	0.02	0.03	0.00	$P108$	0.04	0.06	0.15	0.07	0.00
$P9$	0.10	0.06	0.07	0.06	0.00	$P59$	0.25	0.33	0.02	0.41	0.00	$P109$	0.40	0.54	0.39	0.39	0.00
$P10$	0.04	0.21	0.10	0.18	0.00	$P60$	0.15	0.23	0.07	0.16	0.00	$P110$	0.17	0.28	0.13	0.45	0.00
$P11$	0.10	0.57	0.22	0.24	0.00	$P61$	0.09	0.21	0.04	0.16	0.00	$P111$	0.14	0.11	0.20	0.20	1.00
$P12$	0.08	0.02	0.11	0.02	1.00	$P62$	0.54	0.46	0.12	0.64	0.00	$P112$	0.26	0.30	0.28	0.14	0.00
$P13$	0.32	0.26	0.22	0.61	0.00	$P63$	0.02	0.17	0.05	0.18	0.00	$P113$	0.12	0.03	0.15	0.11	0.00
$P14$	0.34	0.11	0.26	0.06	0.00	$P64$	0.12	0.31	0.03	0.21	0.00	$P114$	0.01	0.02	0.04	0.00	1.00
$P15$	0.14	0.04	0.09	0.10	0.00	$P65$	0.03	0.13	0.06	0.05	0.00	$P115$	0.04	0.02	0.03	0.00	0.00
$P16$	0.17	0.23	0.19	0.28	0.00	$P66$	0.10	0.09	0.20	0.12	0.00	$P116$	0.10	0.15	0.05	0.08	0.00
$P17$	0.21	0.09	0.16	0.07	0.00	$P67$	0.33	0.06	0.20	0.17	0.00	$P117$	0.18	0.03	0.08	0.05	0.00
$P18$	0.31	0.08	0.20	0.09	1.00	$P68$	0.47	0.61	0.19	0.38	1.00	$P118$	0.16	0.37	0.26	0.25	0.00
$P19$	0.17	0.05	0.21	0.03	1.00	$P69$	0.01	0.02	0.09	0.01	0.00	$P119$	0.12	0.10	0.04	0.03	1.00
$P20$	0.02	0.01	0.06	0.02	0.00	$P70$	0.05	0.20	0.27	0.12	0.00	$P120$	0.05	0.05	0.17	0.05	0.00
$P21$	0.25	0.36	0.23	0.10	0.00	$P71$	0.26	0.08	0.47	0.11	1.00	$P121$	0.01	0.00	0.02	0.01	0.00
$P22$	0.05	0.03	0.07	0.07	0.00	$P72$	0.03	0.15	0.24	0.12	0.00	$P122$	0.05	0.01	0.13	0.01	0.00
$P23$	0.05	0.03	0.19	0.00	0.00	$P73$	0.16	0.07	0.10	0.14	0.00	$P123$	0.48	0.38	0.43	0.63	0.00
$P24$	0.24	0.45	0.29	0.78	0.00	$P74$	0.26	0.05	0.14	0.05	0.00	$P124$	0.01	0.00	0.05	0.05	0.00
$P25$	0.15	0.05	0.19	0.07	0.00	$P75$	0.21	0.11	0.05	0.31	0.00	$P125$	0.18	0.26	0.15	0.07	0.00
$P26$	0.14	0.25	0.16	0.03	0.00	$P76$	0.03	0.10	0.10	0.05	0.00	$P126$	0.14	0.02	0.01	0.01	0.00
$P27$	0.05	0.02	0.07	0.08	0.00	$P77$	0.07	0.08	0.42	0.14	1.00	$P127$	0.29	0.12	0.20	0.15	0.00
$P28$	0.07	0.25	0.20	0.09	0.00	$P78$	0.06	0.32	0.02	0.18	0.00	$P128$	0.53	0.58	0.63	0.65	0.00
$P29$	0.45	0.10	0.36	0.12	0.00	$P79$	0.32	0.09	0.28	0.10	0.00	$P129$	0.61	0.19	0.29	0.31	0.00
$P30$	0.19	0.13	0.26	0.19	0.00	$P80$	0.03	0.01	0.13	0.01	0.00	$P130$	0.23	0.22	0.25	0.39	1.00
$P31$	0.42	0.58	0.26	0.46	0.00	$P81$	0.02	0.09	0.02	0.34	0.00	$P131$	0.48	0.62	0.70	0.87	0.00
$P32$	0.14	0.37	0.37	0.15	1.00	$P82$	0.06	0.04	0.13	0.09	0.00	$P132$	0.21	0.29	0.12	0.14	1.00
$P33$	0.33	0.12	0.20	0.39	0.00	$P83$	0.02	0.04	0.02	0.01	0.00	$P133$	0.51	0.68	0.52	0.81	0.25
$P34$	0.16	0.16	0.20	0.04	0.00	$P84$	0.23	0.20	0.21	0.14	0.00	$P134$	0.40	0.32	0.17	0.21	1.00
$P35$	0.14	0.20	0.19	0.44	0.00	$P85$	0.47	0.44	0.20	0.43	0.00	$P135$	0.11	0.06	0.18	0.24	0.25
$P36$	0.09	0.08	0.09	0.08	0.00	$P86$	0.08	0.16	0.17	0.03	0.00	$P136$	0.29	0.19	0.12	0.02	0.25
$P37$	0.20	0.47	0.45	0.66	1.00	$P87$	0.30	0.26	0.13	0.39	0.00	$P137$	0.23	0.11	0.15	0.15	0.00
$P38$	0.00	0.02	0.03	0.00	0.00	$P88$	0.07	0.10	0.05	0.03	0.00	$P138$	0.31	0.54	0.38	0.28	1.00
$P39$	0.13	0.11	0.22	0.15	0.00	$P89$	0.01	0.05	0.28	0.07	1.00	$P139$	0.58	0.69	0.56	0.62	0.25
$P40$	0.31	0.43	0.25	0.17	0.00	$P90$	0.28	0.12	0.16	0.24	0.00	$P140$	0.35	0.54	0.24	0.45	0.00
$P41$	0.01	0.00	0.00	0.01	0.00	$P91$	0.12	0.12	0.52	0.18	1.00	$P141$	0.02	0.07	0.03	0.01	0.00
$P42$	0.13	0.02	0.07	0.10	0.00	$P92$	0.00	0.06	0.15	0.06	0.00	$P142$	0.05	0.04	0.17	0.02	0.25
$P43$	0.16	0.28	0.20	0.26	1.00	$P93$	0.38	0.04	0.05	0.12	0.00	$P143$	0.00	0.00	0.03	0.01	0.00
$P44$	0.01	0.01	0.01	0.02	0.00	$P94$	0.52	0.44	0.12	0.36	0.00	$P144$	0.18	0.15	0.26	0.15	1.00
$P45$	0.00	0.00	0.02	0.00	0.00	$P95$	0.18	0.06	0.22	0.04	0.00	$P145$	0.00	0.00	0.02	0.00	0.00
$P46$	0.20	0.12	0.08	0.03	0.00	$P96$	0.15	0.18	0.50	0.21	1.00	$P146$	0.21	0.02	0.09	0.11	0.00
$P47$	0.23	0.14	0.35	0.32	0.00	$P97$	0.23	0.25	0.10	0.16	0.00	$P147$	0.45	0.38	0.51	0.15	0.00
$P48$	0.21	0.01	0.03	0.04	0.00	$P98$	0.41	0.46	0.55	0.62	1.00	$P148$	0.01	0.05	0.24	0.20	0.00
$P49$	0.19	0.08	0.13	0.09	0.00	$P99$	0.34	0.23	0.19	0.25	0.00	$P149$	0.21	0.07	0.25	0.22	1.00
$P50$	0.08	0.21	0.12	0.18	0.00	$P100$	0.19	0.15	0.25	0.16	0.00						

> $(P(x_{i29} = 1 \mid \tau_i = 3) = 0.36)$, 'Busy, dynamical family' $(P(x_{i47} = 1 \mid \tau_i = 3) = 0.35)$, 'Business-man/-woman' $(P(x_{i77} = 1 \mid \tau_i = 3) = 0.42)$, 'Build a successful career' $(P(x_{i122} = 1 \mid \tau_i = 3) = 0.13)$ and 'Success in life' $(P(x_{i148} = 1 \mid \tau_i = 3) = 0.24)$ has higher cluster specific probabilities for Cluster 3, which corresponds with the description of this main motivational cluster in Section 2.2. Likewise, the items 'Adventurous' $(P(x_{i15} = 1 \mid \tau_i = 1) = 0.14)$, 'Single' $(P(x_{i46} = 1 \mid \tau_i = 1) = 0.20)$, 'Artist' $(P(x_{i93} = 1 \mid \tau_i = 1) = 0.38)$, 'Make dreams come through' $(P(x_{i117} = 1 \mid \tau_i = 1) = 0.18)$ and 'Being unique, different' $(P(x_{i146} = 1 \mid \tau_i = 1) = 0.21)$ has higher cluster specific probabilities for Cluster 1, which corresponds with the description of the main motivational cluster that can be described with the word 'Vitality' in Section 2.2. The items 'Ordinary' $(P(x_{i11} = 1 \mid \tau_i = 2) = 0.57)$, 'Quiet family' $(P(x_{i54} = 1 \mid \tau_i = 2) = 0.37)$, 'Full time housewife' $(P(x_{i78} = 1 \mid \tau_i = 2) = 0.32)$, 'Relaxing at home' $(P(x_{i109} = 1 \mid \tau_i = 2) = 0.54)$ and 'Privacy, tranquility' $(P(x_{i138} = 1 \mid \tau_i = 2) = 0.54)$ has higher cluster specific probabilities for Cluster 2, which corresponds with the description of the main motivational cluster that can be described with the word 'Security' in Section 2.2. And the items 'Cozy' $(P(x_{i24} = 1 \mid \tau_i = 4) = 0.78)$, 'Warm family' $(P(x_{i51} = 1 \mid \tau_i = 4) = 0.47)$, 'Nurse' $(P(x_{i59} = 1 \mid \tau_i = 4) = 0.41)$, 'A sociable evening with friends' $(P(x_{i123} = 1 \mid \tau_i = 4) = 0.63)$ and 'Friendship' $(P(x_{i133} = 1 \mid \tau_i = 4) = 0.81)$ has higher cluster specific probabilities for Cluster 4, which corresponds with the description of the main motivational cluster that can be described with the word 'Harmony' in Section 2.2.
>
> It turns out that running the data set at hand with the specified model renders five clusters in which four of these clusters can be used for marketing purposes. In these four clusters the four main motivational clusters, as described in Section 2.2, are identified. In these four clusters marketeers can recognize distinct groups of respondents, for which differentiated marketing strategies can be made.

Although this solution with five clusters seems practically useful, the posterior predictive p-value for this cluster solution is 0.000. In other words, this cluster solution seems not to be able to reconstruct the data. The reason for using the word "seems" in the previous sentence can be explained after an inspection of the observed $(N(x_j = v, x_{j'} = w))$ and expected $(E(x_j = v, x_{j'} = w \mid \boldsymbol{\pi}, \boldsymbol{\lambda}(.), \boldsymbol{\omega}))$ number of respondents that are used in (2.14).

Table 2.8 displays the observed and expected number of respondents for some item pairs, based on parameters from Gibbs iterations $z = 101, \ldots, 1100$ for $Q_{max} = 5$ of the hierarchical algorithm. Note that the observed numbers are independent

Table 2.8: Observed and Expected Number of Respondents for Item Pairs for Observed Data Set. (Based on the Average Parameters from Gibbs Iteration $z = 101, \ldots, 1100$ for $Q_{max} = 5$ of the Hierarchical Algorithm.)

Capable	Manager	obs.	exp.
0	0	1522.0	1483.2
1	1	74.0	58.5
0	1	527.0	553.2
1	0	55.0	83.1
A little bit shy	Ordinary	obs.	exp.
0	0	1286.0	1280.6
1	1	151.0	155.5
0	1	597.0	594.8
1	0	186.0	189.1
Not suited for family life	Cars/ motor bikes	obs.	exp.
0	0	1928.0	1904.2
1	1	7.0	6.9
0	1	196.0	206.5
1	0	32.0	45.3
No occupation	Friendship	obs.	exp.
0	0	760.0	755.3
1	1	107.0	111.1
0	1	1241.0	1237.5
1	0	65.0	69.1
Helpful	Nurse	obs.	exp.
0	0	1033.0	1004.2
1	1	277.0	265.7
0	1	212.0	240.3
1	0	656.0	667.8

of z. Note furthermore that the expected numbers displayed in Table 2.8 are the averages of the expected numbers computed for iterations $z = 101, \ldots, 1100$.

Based on the examination of the observed and expected number of respondent and given the total number of respondents in the data set, this model is believed to be acceptable. For example the observed and expected number of respondents for item pair 'A little bit shy' and 'Ordinary'. The difference between the observed number of respondents ($N(x_1 = 1, x_{11'} = 1) = 151.0$) and the averaged (over $z = 101, \ldots, 1100$) expected number of respondents ($E(x_1 = 1, x_{11'} = 1 \mid \boldsymbol{\pi}, \boldsymbol{\lambda}(\textbf{.}), \boldsymbol{\omega}) = 155.5$) that pick both items, is relatively small taking the total number of respondents (in this item pair the total number of respondents without missing values is 2220) into account. Also the difference between the observed number of respondents ($N(x_{38} = 1, x_{120'} = 1) = 7.0$) and expected number of respondents ($E(x_{38} = 1, x_{120'} = 1 \mid \boldsymbol{\pi}, \boldsymbol{\lambda}(\textbf{.}), \boldsymbol{\omega}) = 6.9$) that pick both item 'Not suited for family life' and 'Cars/ motor bikes', is relatively small taking the total number of respondents (in this item pair the total number of respondents without missing values is 2163) into account. These relatively small difference between the observed and expected numbers can not only be seen in Table 2.8, but also when examining all other item pairs and other interactions.

Although the posterior predictive p-value indicates that the cluster solution is not able to reconstruct the data, we believe that the cluster solution is relevant. This is based on the relatively small differences between the observed and expected numbers of respondents for all item pairs. Furthermore, the fact that the remaining four clusters from the $Q_{max} = 5$ solution can be identified as the four main motivational clusters, which are practically useful in marketing, does support this beliefs. Also MacLachlan and Mulhern (2004) acknowledge this interaction between statistics and marketing: *'in any empirical problem, the researcher must necessarily use a substantial dose of subjectivity and domain knowledge. This can be aided by computation of some statistical indicators, but ultimately the decision regarding the number of clusters to use in any particular problem will be the result of viewing those indicators in the light of the marketing decision problem at hand.'*.

With the clustering algorithm proposed in Section 2.3 and 2.4 our purpose is to identify motivational clusters of respondents that have more or less the same psychographic description. More specifically, the purpose is to identify the four main motivational clusters that have more of less the same psychographic descriptions as in Section 2.2. Above results show that our proposed algorithm, using prior knowledge about the underlying theoretical framework of Brand Strategy Research (BSR), renders five clear and distinctive latent clusters. Due to the cluster weights of one latent cluster only four clusters from the cluster solution can be used for marketing purposes. These four remaining clusters can be identified as the four main motivational clusters as described in Section 2.2. These four main motivational clusters are recognized by marketeers (Brethouwer et al., 1995, p. 8; Callebaut et al., 1999, p. 55-60), for which differentiated marketing strategies can be made. The results also show that the underlying theory of Brand Strategy Research is validated. By using the prior knowledge about the underlying theoretical framework of BSR, good decisions could be made regarding the inclusion or exclusion of two-way interactions in the analyzing model. Resulting in a cluster solution with the desired four main motivational clusters that are commonly used in day-to-day business by The SmartAgent Company.

Another thing we want to mention (and not go into detail) is that running the clustering algorithm with the data set at hand and using a locally independent model, thus only the 149 main effects, renders a cluster solution with $Q_{max} = 35$ clusters. When interpreting this cluster solution some of these 35 clusters in the model assuming local item independence within the clusters have more of less the same description as the clusters from the model assuming local item dependence within the clusters and some of these 35 clusters are additional clusters that are considered to be artifacts of the use of a cluster model assuming local item independence. Once again, using a model assuming local within cluster independence, instead of using a model assuming local within cluster dependencies, gives a big difference in terms of number of clusters, cluster description and cluster weights.

2.6 Discussion

This chapter presented a Bayesian model based clustering approach that can handle missing values, large data sets and local within cluster dependencies. A hierarchical algorithm was presented that can be used to estimate the number of clusters in the data set and renders estimates of the parameters of the mixture of log-linear models. A pseudo-likelihood ratio test was introduced that is not affected by the fact that the number of possible response vectors (for the data set at hand 2^{149}) is by far out-weighted by the number of observed response vectors. In the example it was illustrated that residuals from this test can be used to determine if and which two-way interactions should added to the model.

The proposed model based clustering approach was applied to a real BSR data set, in which 2294 respondents responded to a list of 149 psychographic items. Using prior knowledge about the underlying theoretical framework of BSR, it was decided which two-way interactions to include in the model. This resulted in a model with 149 main effects and 90 interaction effects. The resulting cluster solution contained $Q_{max} = 5$ clear and distinctive latent clusters. Using the same data set it was also shown that not taking within cluster dependencies into account, that is a model assuming local item independence within the clusters, results in a solution with $Q_{max} = 35$ clusters. This solution is much harder (if not impossible) to interpret than the five cluster solution obtained using a model with within cluster item dependencies.

Appendix I

The BSR Questionnaire
When you think of a particular car, you may also think of the typical person driving such a make or model. Now, when you think of housing* and of the way in which you live (or would like to live), what kind of people go with that? How do these people live, what are their character traits, their hobbies, occupations, et cetera?

Q1. Which character traits fit the best for the person that has the same opinion about housing as you do?

○ a little bit shy
○ adventurous
○ capable
○ cozy
○ energetic
○ a little bit imprudent
○ honest
○ jovial
○ opinionated
○ self-assured
○ serious
○ spontaneous
○ a little impatient
○ assertive
○ cheerful
○ critical
○ enthusiastic
○ gentle
○ intelligent
○ sympathetic
○ ordinary
○ self-confident
○ down-to-earth
○ strong character
○ easygoing
○ balanced
○ classy
○ deliberate
○ leader
○ helpful
○ interested in others
○ neat
○ passionate
○ serene
○ commercial

Q2. Which family or household types fit the best for the person that has the same opinion about housing as you do?

○ a family where everyone goes their own way
○ bachelor
○ busy dynamical family
○ happy family
○ ideal family
○ not suited for family life
○ quiet family
○ single
○ stable family
○ striving for a family
○ artistic household
○ broad-minded family
○ cozy old-fashioned family
○ harmonious family
○ isolated family
○ perfect family
○ rigid family
○ sportive family
○ aristocratic household
○ warm family

* ”housing” may be replaced by the subject of the study, for example, energy, financial services, insurance, health care, et cetera.

Q3. Which occupations fit the best for the person that has the same opinion about housing as you do? Occupations can be done both by males or females.

○ account manager
○ member of the board
○ commercial assistant
○ e-business
○ free-lancer
○ journalist
○ no occupation
○ photographer
○ programmer
○ receptionist
○ shop assistant
○ sports teacher
○ temporary employee
○ vets assistant
○ activity guide
○ business-man/-woman
○ commissioner
○ entrepreneur
○ full time house wife
○ male nurse
○ nurse
○ artist
○ project manager
○ scientist
○ shopkeeper
○ student
○ truck driver
○ volunteer
○ beauty specialist
○ social worker
○ designer
○ financial planner
○ house-husband
○ manager
○ part time house-wife
○ anchor man
○ public servant
○ secretary
○ social worker
○ stylist
○ unemployed

Q4. Which hobbies, interests and/or leisure activities fit the best for the person that has the same opinion about housing as you do?

○ a sociable evening with friends
○ top-notch achievement
○ build a successful career
○ classy parties
○ do odd jobs around the house
○ going to a discotheque
○ make dreams come through!
○ playing chess
○ snow boarding
○ visiting friends and relatives
○ watching TV
○ active sports
○ astrology
○ camping
○ a day out
○ gardening
○ golf
○ religious matters
○ reading magazines
○ working out
○ team sports
○ adventurous holidays
○ being at home quietly
○ cars / motor bikes
○ dine out together
○ going out together
○ investing in stocks
○ swimming
○ shopping
○ surfing the Internet
○ visiting a pub

Q5. Which values fit the best for the person that has the same opinion about housing as you do?

○ anonymity
○ enthusiasm
○ heroism, glory
○ passion
○ recognition of performances
○ self-belief
○ social harmony
○ success in life
○ challenge, stimulation
○ expression, uniqueness
○ independence
○ privacy, tranquility
○ respect
○ self-expression, growth
○ solidarity
○ enjoyable life
○ friendship
○ intimacy
○ rationalism
○ security
○ social alliance
○ status

Chapter 3

Reducing the Optimal to a Useful Number of Clusters for Model Based Clustering

Abstract

Market segmentation is the process in marketing of grouping customers into smaller subgroups, according to a certain segmentation basis. Market segmentation is only practically useful if the effectiveness and profitability of marketing activities are influenced substantially by discerning separate homogeneous groups of customers. Using six criteria, described by Wedel and Kamakura (2000), the effectiveness and profitability of market segmentation can be determined. However, using model based clustering techniques to group customers, the statistically optimal solution often contains too many clusters for the intended marketing purposes. Using the six criteria of good market segmentation, an information criterion and two conjectures, describing the geometry of model based clustering models, presented by Hoijtink and Notenboom (2004), a procedure to reduce the statistically optimal number of clusters to a smaller number, suited for the intended marketing purposes, is presented.

3.1 Introduction

Many people today believe that market segmentation, that is *the grouping of customers who share common aspects*, is the key strategic concept in marketing (Elmore-Yalch, 1998; MacLachlan and Mulhern, 2004; Verhage and Cunningham, 1984, p.186). Marketeers perform market segmentation, expecting to find some segments or clusters in a particular market, responding differently than others to (relevant) marketing items. These marketing clusters can be used for several (differentiated) marketing actions, like for example, differentiated mail packages, differentiated catalogues, prevention against churn in specific clusters, cross/up-selling strategies, et cetera.

From a market segmentation study in the domain housing it is known, that the (statistically) optimal number of market segments (or clusters) in the Dutch housing market is 35. Of course, it is nice to known that in The Netherlands the total housing market can be divided into 35 clusters, but as money and time is limited in most marketing companies, it is undesirable to interpret, describe or make differentiated marketing plans for these 35 market segments. In marketing there is a wish of dividing a market as effectively as possible. Resulting in a market segmentation with as few clusters as possible to fully describe the total market and, finally, to solve the marketing decision problem at hand. From this point of view it is clear, that it is possible that there may be a difference between the statistically optimal solution and the solution suited for the intended marketing purposes. This chapter describes an algorithm to reduce the (statistically) optimal number of clusters to a smaller number suited for marketing purposes. The effectiveness and profitability of these market segments or clusters are determined using six criteria, described by Wedel and Kamakura (2000, p.4-5).

The structure of this chapter is as follows. Section 3.2 briefly describes market segmentation. Section 3.3 introduces the clustering algorithm that is used for market segmentation in this chapter. In Section 3.4 the procedure to reduce the (statistically) optimal number of clusters to a smaller number, suited for marketing purposes, is described. Section 3.2, 3.3 and 3.4 are illustrated using a market segmentation study in the domain housing. Section 3.5 introduces a marketing application concerning a Dutch mail order company. This chapter concludes with a discussion in Section 3.6.

3.2 Market Segmentation

3.2.1 Introduction

Market segmentation is an essential element of marketing in industrialized countries. Goods can no longer be produced and sold without considering customer needs and recognizing the heterogeneity of those needs (Wedel and Kamakura, 2000, p.3). Research in the private sector (Elmore-Yalch, 1998) has shown conclusively, that if you can find segments or clusters that, (1) you can identify and differentiate, (2) will remain effectively stable, and (3) can effectively be reached, a company can increase sales and profits by marketing to these segments or clusters, beyond profits possible from treating the market as homogeneous. The concept of market segmentation is further described in Section 3.2.2. Section 3.2.3 introduces a market segmentation study in the domain housing.

3.2.2 Using Segmentation in Marketing

The basic problem of market segmentation is *the grouping of customers who share common aspects*. In his article Smith (1956) stated that it was better to recognize several customer demand schedules. As Wedel and Kamakura (2000, p.3) mention in their book, Smith recognized the existence of heterogeneity in the demand of goods and services. Smith stated: *'Market segmentation involves viewing a heterogeneous market as a number of smaller homogeneous markets, in response to differing preferences, attributable to the desires of customers for more precise satisfaction of their varying wants.'* The idea of market homogeneity in marketing theory is also rejected by Alderson (1965, p.2), whose theory of marketing was also based on the concept of heterogeneity. Heterogeneity in its most extreme case, also called complete market heterogeneity (Bell and Vincze, 1988, p.290), means that each customer is unique in at least one important aspect (or, in Smith's terminology demand schedule) and in this aspect is like no other person. In other aspects customers may be more or less similar. The concept of complete market heterogeneity overstates the reality of practical marketing. Pockets of similarity, known as market segments or clusters, do exist. Thus, in spite of the fact that each customer differs from every other, it is still true that each customer tends to be more like some customers than like others (Elmore-Yalch, 1998).

Marketeers do not attribute these similarities to chance. They know that there are basic differences among market segments. Marketing practice requires that a marketeer knows how market segments differ in their attitudes and susceptibilities to marketing efforts. From this knowledge, separated or differentiated marketing strategies can be made (Bell and Vincze, 1988). For example, differentiated mail packages, differentiated catalogues, prevention against churn in specific clusters, cross/up-selling strategies, et cetera.

However, market segmentation is only practically useful if the effectiveness and profitability of such marketing activities are influenced substantially by discerning separate homogeneous groups of customers. Using six criteria, described by Wedel and Kamakura (2000, p.4-5), the effectiveness and profitability of market segmentation can be determined. Below, brief descriptions of these six criteria are given:

1. Identifiability: a cluster must be clearly defined. It must be clear who is in the cluster.

2. Substantiality: a cluster must be large enough to ensure the profitability of developing a differentiated marketing strategy.

3. Accessability: a cluster must be reachable through promotional or distributional marketing activities.

4. Responsiveness: a cluster must respond uniquely to marketing activities.

5. Stability: a cluster must be stable in time, at least for a period long enough for identification of the clusters, implementation of a differentiated marketing strategy and to produce profitable results.

6. Actionability: a cluster and the differentiated marketing strategy must be consistent with the goals and core competencies of the company.

3.2.3 Application: the Dutch Housing Market

Competition in the Dutch housing market is fierce. Due to growing working mobility, electronic access possibilities and a growing (inter)national orientation, the customer is not only focused on the local housing market. More than ever, project developers, investors, real estate consultants and governmental housing institutions must be aware of their target group. Who are they? What drives them? What do they want? How can I reach them? Or, like the philosopher Heidegger (1991) mentioned: '*only when we know how to live, we can build.*'; it is important to understand what the key drivers of customers are on the housing market. In order to find out what kind of housing customers there are in The Netherlands, a market segmentation study is conducted. In this segmentation study the framework of

Brand Strategy Research (BSR)* (Brethouwer et al., 1995, p.8; Oppenhuisen 2000, p.79-81) is used.

The data that are used in this chapter are collected by The SmartAgent Company, The Netherlands (`www.smartagent.nl`) through a questionnaire on the internet in 2000. The process behind BSR is the following: a respondent is asked to characterize a person, resembling himself, who looks or feels the same as the respondent towards housing. The whole BSR questionnaire consists of five questions, each containing multiple psychographic items. The first question contains items that describe a person's character. The second question tells something about a person's type of household. The third gives a person's occupations, the fourth question tells something about a person's hobbies and interests and the last questions tells which values a person can have in life. Appendix I displays the BSR questions. In total there are 149 psychographic items answered by 2294 respondents. For each question a respondent has to pick the items which describe the person he has in mind the best. Because each question contains a broad range of items, it is unlikely that a respondent can not pick an item from the item list.

Using a model based clustering algorithm, which is described in the next section, the 2294 respondents are clustered according to the 149 psychographic items. Resulting in groups of customers who have (more or less) the same view, motivations and attitude with respect to housing. In order to further describe these motivational clusters found, the online questionnaire also contains observable items, like for example, demographical items (i.e., gender, age, education, marital status, et cetera), economical items (i.e., working position, social economic status, prosperity, income, et cetera), housing specific items (i.e., preferred house, preferred neighborhood, preferred price, et cetera), et cetera. What the different types of customers are in the Dutch housing market and how they can be described using the observable items, is shown in Section 3.4.2.

3.3 Model Based Clustering Techniques

3.3.1 Introduction

Much of the literature about market segmentation has evolved around the technique of identifying clusters from data. See Wedel and Kamakura (2000, Chapter 3)

*BSR is based on Adler's social-psychology theory (Callebaut et al. 1999, p.55-60) and provides a framework for understanding customers at the 'deepest' level. This motivational level gives knowledge of customer's fears, beliefs and values, thus providing an understanding of the fundamental motivations that drive (future) purchase decisions of customers. The interested reader is referred to `www.smartagent.nl` for more information about BSR.

for an extensive overview of these techniques. A substantial part of this literature are comparative papers, that contrast the most widely used clustering techniques. See MacLachlan and Mulhern (2004) for an overview of these papers. More recent papers (MacLachlan and Mulhern, 2004; Magidson and Vermunt, 2002; Mulhern and MacLachlan, 2003) compare mixture models (mixture models, model based clustering algorithms and latent class models are all models coming from the generalized latent variable model framework) with more traditional cluster techniques, like K-means, et cetera. Within the context of market segmentation a number of papers do suggest better market segments, when using mixture models. See for an overview of these papers MacLachlan and Mulhern (2004). An important advantage of mixture modelling over traditional clustering techniques is the statistical framework mixture models are based on. A disadvantage is, that mixture models are less available in popular statistical software than the traditional clustering models. This often results in researchers making their own software (Hoijtink and Notenboom, 2004; Ter Braak et al., 2003; Van Hattum and Hoijtink, 2009b; Wedel and Kamakura, 2000, Chapter 6) and commercial packages, like for example: Glimmix (Wedel and Kamakura, 2000, p.181-186) and LatentGold (Vermunt and Magidson, 2000). For other advantages and disadvantages of mixture modelling, see MacLachlan and Mulhern (2004).

3.3.2 Model Based Clustering Algorithm

This chapter uses a model based clustering approach, that is proposed in Van Hattum and Hoijtink (2009b). The reason for choosing this clustering technique is, that it can be applied to the large (in terms of number of items and customers) data sets, that are often encountered in marketing and can handle missing values. The proposed clustering technique has incorporated a missing value mechanism that can deal with missing values.

The core of the model based clustering algorithm is the estimation of a set unknown parameters, that are:

- For all J items assuming within cluster independence of the item responses: within each cluster $q = 1, \dots, Q$, a vector $\boldsymbol{\pi}_q$, containing the cluster specific probabilities of picking the items, is estimated. Note that $\boldsymbol{\pi}_q = \{\pi_{1|q}, \dots, \pi_{j|q}, \dots, \pi_{J|q}\}$ and $\pi_{j|q}$ is the probability of picking item j in cluster q;

- A vector $\boldsymbol{\omega} = \{\omega_1, \dots, \omega_q, \dots, \omega_Q\}$, containing the cluster weights, that is the proportion of customers allocated to each cluster, is estimated;

Table 3.1: Example: Data set and Unobserved Cluster Memberships ($\boldsymbol{\tau}$).

Customer	Gentle	Honest	Capable	Leader	$\boldsymbol{\tau}$
1	1	1	0	0	1
2	1	0	0	0	1
3	1	1	1	0	1
4	1	1	0	1	1
5	0	0	1	1	2
6	0	1	1	1	2
7	0	0	1	1	2

- A vector $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_i, \ldots, \tau_N\}$, containing the unobserved cluster memberships for each of the N customers, is estimated. Note that for customer $i = 1, \ldots, N$, $\tau_i \in \{1, \ldots, Q\}$.

The interested reader is referred to Van Hattum and Hoijtink (2009b) for the technical details of the proposed model based clustering algorithm. In this chapter the framework of the model based clustering algorithm is illustrated using a simple data set (for simplification only $J = 1, \ldots, 4$ BSR items and $N = 1, \ldots, 7$ customers are used) in Table 3.1 and cluster weights and cluster specific item probabilities in Table 3.2. From the data set in Table 3.1 it is clear that the data set can be divided into two groups of customers, with (more or less) the same answer pattern on the four items. As a result of the division each customer is allocated to one of the two clusters, that is cluster 1 or cluster 2. This is shown in column $\boldsymbol{\tau}$ in Table 3.1, that are the unobserved cluster memberships. As can be seen from the unobserved cluster memberships, the proportion of customers allocated to cluster 1, is four out of seven customers and to cluster 2 is three out of seven customers. These proportions are shown in row $\boldsymbol{\omega}$ in Table 3.2, that are the cluster weights. Furthermore, the cluster specific probabilities of picking the items are calculated using the unobserved cluster memberships. The cluster specific item probabilities for this example are shown in Table 3.2. For example, three out of four customers, allocated to cluster 1, picked item 'Gentle', resulting in $\pi_{2|1} = 0.75$, that is the probability of picking item 'Gentle' ($j = 2$) in cluster 1. Likewise, the probability of picking item 'Gentle' ($j = 2$) in cluster 2 is 0.33, et cetera.

According to two conjectures (Hoijtink and Notenboom, 2004) the model based clustering algorithm renders the globally optimal solution. The main implication of these two conjectures is that the globally optimal solution has Q_{max} latent clusters and all solutions with $Q < Q_{max}$ latent clusters are known to be locally optimal

Table 3.2: Example: Cluster Weights ($\boldsymbol{\omega}$) and Cluster Specific Item Probabilities ($\boldsymbol{\pi}$).

Item	Cluster 1	Cluster 2
$\boldsymbol{\omega}$	0.57	0.43
Gentle	$\pi_{1\|1} = 1.00$	$\pi_{1\|2} = 0.00$
Honest	$\pi_{2\|1} = 0.75$	$\pi_{2\|2} = 0.33$
Capable	$\pi_{3\|1} = 0.25$	$\pi_{3\|2} = 1.00$
Leader	$\pi_{4\|1} = 0.25$	$\pi_{4\|2} = 1.00$

solutions, that are non-overlapping combinations of the Q_{max} clusters.

Running the model based clustering algorithm with the data set concerning the Dutch housing market, as described in Section 3.2.3, renders a globally optimal solution with $Q_{max} = 35$ latent clusters. Or, in other words, the total Dutch housing market can be divided into 35 groups of customers who have (more or less) the same view, motivations and attitude with respect to housing. Of course, it is nice to known that in The Netherlands the total housing market can be divided into 35 clusters, but as time and money is limited in most marketing companies, it is undesirable to interpret, describe or make differentiated marketing plans for these 35 market segments, taking the six criteria of good market segmentation into account. In marketing there is a wish of dividing a market as effectively as possible. Resulting in a market segmentation with as few clusters as possible to fully describe the market and, finally, to solve the marketing decision at hand. Section 3.4.2 describes how these 35 clusters are reduced to a smaller number, suited for the intended marketing purposes. Point of departure for the reduction algorithm is the main implication of the two conjectures, that is the globally optimal solution has Q_{max} latent clusters and all solutions with $Q < Q_{max}$ latent clusters are known to be locally optimal solutions. The model based clustering approach, proposed by Van Hattum and Hoijtink (2009b), is elaborated such that clusters are not only separated, but also combined.

3.4 Using Information Criteria in a Decision Tree

3.4.1 Reduction Algorithm

An important reason for choosing the model based clustering algorithm, described in Section 3.3.2, is because it is based on two conjectures with respect to the geometry of model based clustering models. The main implication of these two conjectures form the point of departure for the reduction of the statistically optimal solution into a solution suited for marketing purposes. The model based clustering algorithm starts with one cluster, containing all customers. Clusters are split and reallocated between clusters, until the cluster solution reaches its globally optimal solution with Q_{max} clusters. All cluster solutions with fewer clusters than this global optimum are known to be locally optimal solutions in which the clusters are non-overlapping combinations of the Q_{max} clusters (Hoijtink and Notenboom 2004).

In order to reduce the number of clusters, the model based clustering algorithm is used in reverse order. Or, using an agglomerative method, the globally optimal solution with Q_{max} clusters is reduced to a solution with $Q_{desired}$ clusters, that is a solution with a desired number of cluster, suited for the intended marketing purposes. This new method combines two clusters from a cluster solution with Q clusters, resulting in a cluster solution with $Q - 1$ clusters. But, with Q clusters there are $\binom{Q}{2}$ possible 'ways' to combine two of the Q clusters, resulting in $\binom{Q}{2}$ different cluster solutions with $Q - 1$ clusters. Because $Q - 1 < Q_{max}$ these $\binom{Q}{2}$ cluster solutions, with $Q - 1$ clusters, are all known to be locally optimal solutions. The above is illustrated using the example in Figure 3.1. In this example $Q_{max} = 4$. Part I of Figure 3.1 shows the four clusters. With $Q_{max} = 4$ there are $\binom{4}{2}$ possible 'ways' to combine two of the Q_{max} clusters in order to get solutions with $Q = 3$ clusters. These solutions with $Q = 3$ clusters are locally optimal solutions, but can be very useful from a marketing perspective. The possible solutions with $Q = 3$ clusters are shown in part II of Figure 3.1. In order to pick the 'best' locally optimal solution from these $\binom{Q}{2}$ cluster solutions, an information criterion (Lin and Dayton, 1997; Vermunt and Magidson, 2000, p.61) is used.

Basically, information criteria impose a penalty on the likelihood that is related to the number of parameters estimated (Kamakura and Wedel, 2000, p.91-92). In general information criteria have the following form:

$$Criterion = -2 \cdot logL + P \cdot d \tag{3.1}$$

Here $logL$ is the log-likelihood function, P is the number of parameters in the model and d is some constant. Where $-2 \cdot logL$ is the measure of fit and $P \cdot d$ is the measure of model complexity. The constant $P \cdot d$ weights the increase in fit against the additional number of parameters estimated. In statistics this penalty has always been a topic of discussion (Akaike, 1987; Andrews and Currim, 2003; Bozdogan, 1987; Wedel and Kamakura, 2000, p.92), resulting in criteria with different penalty functions. Some well known information criteria are: Akaike Information Criterion (AIC, where $d = 2$) (Akaike, 1987), Consistent Akaike Information Criterion (CAIC,

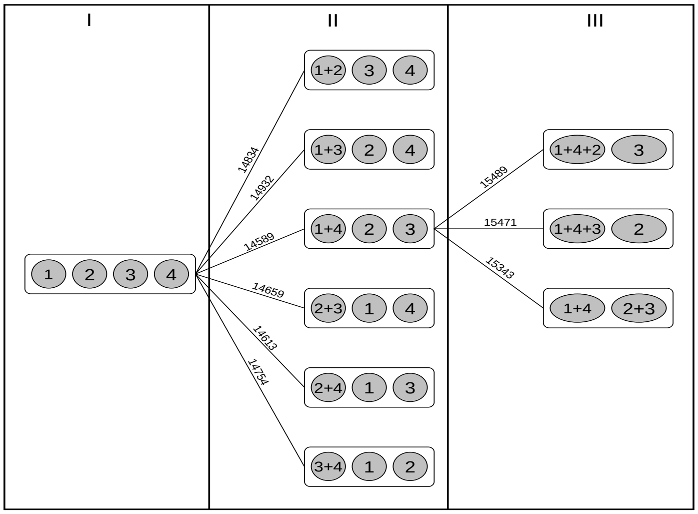


Figure 3.1: Decision Tree

where $d = ln(N+1)$) (Bozdogan, 1987; Wedel and Kamakura, 2000, p.92), Bayesian Information Criterion (BIC, where $d = ln(N)$) (Congdon, 2005, pp.472-474; Wedel and Kamakura, 2000, p.92), et cetera. Since information criteria can be seen as the distance between the current model and the true model, the model with the lowest value for the information criteria is most preferred.

But despite of all the discussion about which information criterion is the best, the choice of which information criteria, or, in other words, the choice of which penalty function, is no point of discussion in this chapter. The reason for that is, when comparing the $\binom{Q}{2}$ cluster solutions with $Q-1$ clusters, the number of parameters, P, and the number of customers, N, are equal for each cluster solution. So, for each of the $\binom{Q}{2}$ cluster solutions with $Q-1$ the same penalty applies. As such this chapter uses $-2 \cdot logL$ as criterion function to decide which of the cluster solutions is the best.

Above described agglomerative method and the use of the $-2 \cdot logL$ criterion in selecting the 'best' locally optimal solution, are incorporated in a decision tree. The root of this decision tree represents the globally optimal solution with Q_{max} clusters and the leaves represent the $\binom{Q}{2}$ locally optimal solutions with $Q-1$ clusters. Using $-2 \cdot logL$ in each branch of the tree, a decision has to be made, resulting in the 'best' locally optimal cluster solution with $Q_{desired}$ clusters.

The above is illustrated using the example tree in Figure 3.1. In this example $Q_{max} = 4$, that is the statistically optimal number of clusters and $Q_{desired} = 2$, that is the desired number of clusters, suited for the intended marketing purposes. The root of the decision tree is shown in part I of Figure 3.1. With $Q_{max} = 4$ there are $\binom{4}{2}$ possible (locally optimal) solutions with $Q = 3$ clusters. The possible solutions with $Q = 3$ clusters are shown in part II of Figure 3.1. Looking at the $-2 \cdot logL$ values which are indicated on each edge, the minimum $-2 \cdot logL$ value is at the $Q = 3$ solution in which cluster 1 and 4 from the $Q_{max} = 4$ solution are combined. This solution is considered to be the 'best' locally optimal solution with $Q = 3$ clusters. From this 'best' solution with $Q = 3$ clusters, the decision step is repeated again. There are $\binom{3}{2}$ possible solutions in which two of the three clusters are combined. These possible (locally optimal) solution are shown in part III of Figure 3.1. From this part it can be seen that the minimum $-2 \cdot logL$ value is at the solution in which cluster 2 and 3 from the $Q = 3$ solution are combined, resulting in the 'best' locally optimal solution with $Q_{desired} = 2$ clusters.

3.4.2 Application

As described in Section 3.3.2, running the model based clustering algorithm with the data set at hand and a cluster model with the 149 BSR items, renders a globally optimal solution with $Q_{max} = 35$ clusters. But, according to the main applications of these conjectures, when the globally optimal solution has $Q_{max} = 35$ latent clusters, all other solutions with $Q < 35$ latent clusters are known to be locally optimal solutions. As was stated earlier in this chapter, time and money are limited in marketing companies and using all these $Q_{max} = 35$ clusters is not desired. So, taken into account the six criteria of good market segmentation, the statistically optimal solution with $Q_{max} = 35$ clusters should be reduced to a solution suited for the intended marketing purposes. This reduction in number of clusters is achieved using the algorithm, described in Section 3.4.1 and evaluated using the six criteria of good marketing segmentation, described in Section 3.2.2.

But what is $Q_{desired}$ in the Dutch housing market? It may be clear that marketeers try to find a cluster solution that is a trade off between the statistically optimal solution (in this application 35 segments) and a solution suited for the intended marketing purposes. Also MacLachlan and Mulhern (2004) acknowledge this interaction between statistics and marketing: *'in any empirical problem, the researcher must necessarily use a substantial dose of subjectivity and domain knowledge. This can be aided by computation of some statistical indicators, but ultimately the decision, regarding the number of clusters to use in any particular problem, will* *be the result of viewing those indicators in the light of the marketing decision problem at hand'.* Besides the substantial dose of subjectivity and domain knowledge of the researcher or another expert, the decision about the desired number of clusters in the Dutch housing market is supported with the help of the six criteria of good market segmentation (as described in Section 3.2.2).

As was mentioned earlier in this chapter, it is undesirable to interpret, describe or making differentiated marketing plans for 35 clusters. A solution with 30 or 25 clusters also takes too much time to describe and interpret. That's why the researcher decides to describe and interpret the reduced solutions with 15 clusters and the solutions with smaller number of clusters. Of course, this is a rather subjective decision, but it is an assessment between domain knowledge, available time and money. In order to determine what $Q_{desired}$ is, the cluster solution with 15 clusters is further reduced, until the most effective and most profitable solution is found.

When interpreting these cluster solutions, the reduced solution with $Q_{desired} = 6$ clusters is the most effective and most profitable one. This is shown using the six criteria of good market segmentation:

1. Identifiability: a cluster must be clearly defined. It must be clear who is in the cluster. Table 3.3 shows for each BSR item (for simplification only the BSR items about a person's character traits and a person's values are shown), the cluster specific probabilities. Using these cluster specific probabilities for the the BSR items and the cluster specific probabilities for the observable items, that are demographical items, economical items, housing items, et cetera, in Table 3.4 (for simplification only a few observable items are shown), each of the six clusters is described. This results in the following cluster descriptions:

    - Cluster 1: persons from this cluster strive for harmony in every aspect of life and harmonious relations with all people they meet in daily life. Harmony between family life and career, between friends and neighbors, relations in general and the rules and values of society. Families with children have a higher probability to occur in this cluster. Persons in this cluster are not ambitious in their career. They are low and moderate educated and have an average income. They think its important to know your neighbors by name. Social cohesion in their neighborhood is an important factor. Also living closely to family and friends is important. Most of the persons in this cluster have never left their birthplace. Persons from this cluster prefer to live near schools, shops and parks/playgrounds, where children can play and people can meet and chat. In general terraced houses can be found in this cluster;

Table 3.3: Cluster Specific Item Probabilities for the $Q_{desired} = 6$ (or Best Local Mode) Solution.

Cluster		1	2	3	4	5	6
Cluster size		28.8%	18.8%	18.1%	5.1%	19.7%	9.5%
BSR question about character traits	A little bit shy	0.08	0.26	0.03	0.08	0.32	0.12
	Easygoing	0.17	0.17	0.15	0.03	0.03	0.34
	A little bit impatient	0.22	0.20	0.33	0.13	0.28	0.14
	Honest	0.58	0.67	0.46	0.39	0.60	0.60
	Assertive	0.12	0.03	0.16	0.27	0.05	0.10
	Critical	0.25	0.13	0.36	0.59	0.34	0.12
	Interested in others	0.57	0.55	0.34	0.38	0.29	0.39
	Gentle	0.16	0.25	0.10	0.05	0.19	0.13
	Jovial	0.07	0.04	0.06	0.03	0.07	0.12
	Neat	0.13	0.24	0.15	0.03	0.12	0.08
	Ordinary	0.25	0.53	0.21	0.02	0.51	0.26
	Capable	0.04	0.03	0.13	0.03	0.06	0.03
	Spontaneous	0.54	0.32	0.19	0.26	0.06	0.42
	Strong character	0.16	0.06	0.30	0.49	0.16	0.18
	Adventurous	0.08	0.02	0.06	0.29	0.06	0.19
	Sympathetic	0.23	0.24	0.15	0.27	0.18	0.25
	Energetic	0.11	0.11	0.20	0.21	0.05	0.16
	Self-confident	0.14	0.09	0.27	0.28	0.14	0.08
	Leader	0.08	0.03	0.31	0.15	0.06	0.13
	Classy	0.03	0.01	0.07	0.09	0.02	0.02
	Serious	0.13	0.35	0.32	0.14	0.38	0.19
	A little impudent	0.06	0.03	0.08	0.09	0.03	0.05
	Commercial	0.02	0.02	0.27	0.00	0.11	0.05
	Cozy	0.59	0.48	0.22	0.34	0.29	0.56
	Self-assured	0.10	0.03	0.17	0.18	0.13	0.16
	Deliberate	0.05	0.19	0.20	0.10	0.36	0.08
	Passionate	0.08	0.01	0.03	0.22	0.01	0.08
	Serene	0.09	0.17	0.23	0.03	0.35	0.11
	Intelligent	0.20	0.06	0.36	0.70	0.23	0.19
	Opinionated	0.17	0.11	0.21	0.21	0.25	0.25
	Helpful	0.51	0.67	0.20	0.11	0.33	0.53
	Down-to-earth	0.18	0.37	0.33	0.05	0.55	0.17
	Enthusiastic	0.36	0.14	0.18	0.35	0.04	0.28
	Balanced	0.10	0.13	0.23	0.15	0.19	0.10
	Cheerful	0.32	0.25	0.12	0.27	0.10	0.38
BSR question about values	Self-belief	0.61	0.61	0.63	0.52	0.57	0.56
	Self-advancement, growth	0.39	0.15	0.36	0.56	0.19	0.18
	Enthusiasm	0.32	0.26	0.26	0.17	0.17	0.33
	Enjoying life	0.75	0.68	0.68	0.54	0.57	0.70
	Social harmony	0.22	0.25	0.11	0.06	0.27	0.12
	Friendship	0.67	0.78	0.46	0.37	0.52	0.82
	Social alliance	0.25	0.32	0.21	0.26	0.23	0.27
	Passion	0.19	0.02	0.12	0.38	0.06	0.28
	Solidarity	0.10	0.13	0.12	0.11	0.25	0.16
	Intimacy	0.16	0.11	0.13	0.27	0.10	0.17
	Privacy, tranquility	0.32	0.52	0.35	0.35	0.62	0.26
	Respect	0.63	0.76	0.57	0.35	0.58	0.63
	Security	0.46	0.68	0.20	0.26	0.38	0.14
	Anonymity	0.01	0.04	0.02	0.02	0.11	0.03
	Rationality	0.04	0.03	0.16	0.05	0.14	0.04
	Status	0.01	0.00	0.03	0.00	0.01	0.04
	Recognition of performances	0.16	0.10	0.31	0.24	0.23	0.20
	Heroism, glory	0.00	0.00	0.02	0.00	0.00	0.03
	Being unique, different	0.10	0.01	0.06	0.43	0.06	0.07
	Independency	0.25	0.35	0.50	0.44	0.55	0.36
	Success in life	0.11	0.05	0.31	0.11	0.08	0.25
	Challenge, stimulation	0.17	0.04	0.30	0.39	0.10	0.30

Table 3.4: Cluster Specific Probabilities for the Observable Items, for the $Q_{desired} = 6$ (or Best Local Mode) Solution.

Cluster		1	2	3	4	5	6
Cluster size		28.8%	18.8%	18.1%	5.1%	19.7%	9.5%
Age	20-29 years	0.08	0.14	0.13	0.22	0.06	0.14
	30-39 years	0.22	0.17	0.26	0.24	0.20	0.29
	40-49 years	0.25	0.21	0.25	0.22	0.24	0.28
	50-59 years	0.23	0.24	0.19	0.17	0.25	0.18
	60-69 years	0.13	0.12	0.11	0.09	0.17	0.08
	70 years and over	0.08	0.12	0.04	0.03	0.08	0.03
Education	High	0.13	0.29	0.46	0.51	0.23	0.24
	Middle	0.66	0.52	0.49	0.44	0.57	0.65
	Low	0.21	0.19	0.05	0.06	0.20	0.11
Type of family	Single family, young	0.01	0.10	0.05	0.10	0.01	0.03
	Single family, old	0.11	0.38	0.10	0.15	0.10	0.08
	Family with children	0.44	0.17	0.39	0.31	0.39	0.50
	Family with adults, young	0.04	0.05	0.11	0.12	0.05	0.09
	Family with adults, old	0.41	0.30	0.35	0.31	0.45	0.30
Social economic status	A -high-	0.11	0.11	0.26	0.24	0.11	0.14
	Bb	0.25	0.24	0.37	0.33	0.29	0.30
	Bo	0.22	0.24	0.22	0.25	0.27	0.24
	C	0.38	0.37	0.14	0.16	0.31	0.30
	D -low-	0.04	0.04	0.01	0.01	0.02	0.01
Income	Lower than average	0.22	0.34	0.09	0.17	0.16	0.15
	Average	0.35	0.31	0.17	0.20	0.32	0.29
	Higher than average	0.43	0.34	0.75	0.62	0.53	0.56
Type of house	Own house	0.67	0.47	0.79	0.67	0.72	0.75
	Rental	0.33	0.53	0.21	0.33	0.28	0.25
Domain specific statements	It is important to live in a dynamic environment	0.01	0.03	0.05	0.10	0.02	0.03
	A neighborhood with the same type of people is important	0.09	0.12	0.16	0.11	0.14	0.09
	Anonymous living is important	0.02	0.06	0.03	0.04	0.04	0.01
	Cozy living is important	0.47	0.37	0.27	0.30	0.35	0.45
	Exclusive housing is important	0.00	0.02	0.06	0.06	0.01	0.02
	Functional living is important	0.10	0.14	0.23	0.25	0.18	0.13
	I prefer to live like a hermit	0.38	0.51	0.32	0.26	0.33	0.25
	It is important to have a lot of social contacts in the neighborhood	0.23	0.19	0.17	0.18	0.23	0.23
	It is important to have many visitors in my house	0.15	0.11	0.11	0.21	0.10	0.21
	It is important to live free, with a lot of space	0.33	0.38	0.51	0.51	0.46	0.35
	It is important to live in a house that is different from others	0.03	0.05	0.10	0.17	0.05	0.07
	It is important to live luxurious	0.03	0.06	0.17	0.12	0.04	0.06
	My house is always open for everybody	0.42	0.31	0.30	0.38	0.32	0.49
	My house is basically a place to sleep	0.01	0.03	0.01	0.02	0.01	0.01
	It is important to have privacy	0.49	0.59	0.57	0.50	0.56	0.43
	It is important to live secure	0.58	0.52	0.53	0.43	0.57	0.53
	It is important to have social cohesion in the neighborhood	0.43	0.33	0.32	0.37	0.36	0.49
	It is important to live solitarily	0.01	0.04	0.03	0.03	0.03	0.01

- Cluster 2: persons from this cluster are mainly oriented on their peer group and the rules and values of this group. Following these rules and orientation on the peer group creates a feeling of security and belonging. Persons in this cluster choose to live in an environment where they can live, work and shop. An important aspect in their neighborhood is privacy and anonymity. Their houses must be a safe place to live in. In general low educated persons with low income have higher probability to be in this cluster. Also the age of these persons is in general higher. This cluster contains in general more single households;

- Cluster 3: persons from this cluster are career oriented and aspire a certain (high) status in life in connection with certain status symbols and conspicuous consumption. This goes along with manifestative behavior and attitudes as well as a need for control. Translated into their housing, these persons prefer to live in large, detached houses, like villa's, bungalows and penthouses. They prefer to live in a neighborhood with their own type of persons. Luxury housing and status objects are important aspects. Social contacts and coziness in their neighborhood are not appreciated. Bonding with their house and neighborhood is low, with in general higher and faster moving rates. Families with children can also be found in this cluster. Compared to the families in Cluster 1, families in this cluster are not as comfy. These families are more dynamic; each family member goes his own way, showing conspicuous consumption wherever possible. Persons in this cluster are in general high educated, with a high income. They have the highest social economic status, with a lot of prosperity. A lot of directors/CEOs can be found in this cluster;

- Cluster 4: persons from this cluster are self conscious and self-confident in their attitude towards (choices in) life and energetic, vital and passionate in their behavior. Persons in this cluster prefer living in an apartment, in a large town or city center. Living in town is interesting, as you can lead your own life. The creative and independent character will certainly be shown in the choice of houses. In general young, well educated persons can be found in this cluster. They are on the eve of a successful career. Also more single households can be found in this cluster;

- Cluster 5: persons from this cluster are, in terms of description, a combination of the persons in cluster 2 and 3. They are always trying to find a good combination between family and career. Persons from this cluster are also oriented on their peer group and the rules and values of this group, but not as strict as persons from cluster 2. These persons are in general middle aged and on the eve of becoming 'empty nesters'. This new phase in life leads to different demands in housing. Bonding with their residential place is due to their work, children or sports, rather than family and friends. In general persons in this cluster are sportive, social and like to be busy with their houses and gardens. Houses that can be found in this cluster are in general (semi)detached and luxury apartments;

- Cluster 6: persons from this cluster are, in terms of description, a combination of the persons in cluster 1 and 4. These persons can best described by the word 'normal'. They are called 'Monsieur Toutlemonde' in France, 'Joe Sixpack' in The USA, 'Otto Normalverbraucher' in Germany or 'Jan Modaal' in The Netherlands. They are the common suburban family man; in general middle educated, middle aged, working as an employee with an average income, part of a family with an average number of children, living in a terraced house in a normal environment, et cetera.

Given these descriptions the six clusters are clearly defined. It is clear who is in the cluster;

2. Substantiality: a cluster must be large enough to ensure the profitability of developing a differentiated marketing strategy. From the row Cluster size in Table 3.3 it can be seen that all six clusters has substantial weights, that is cluster 1 = 28.8%, cluster 2 = 18.8%, cluster 3 = 18.1%, cluster 4 = 5.1%, cluster 5 = 19.7% and cluster 6 = 9.5%. Whether these substantial clusters are profitable enough, can be decided after a differentiated marketing strategy has taken place (Stanton and Pires, 1999). However, the descriptions of the six clusters provide relevant information on how to communicate with them and to set-up all kind of (differentiated) marketing strategies;

3. Accessability: a cluster must be reachable through promotional or distributional marketing activities. According to the literature this criterion is less appealing in psychographic segmentation (Wedel and Kamakura, 2000, p.16). However, in above described motivational segmentation study, demographical, economical and housing specific items are used in order to further describe the six motivational clusters. Using these observable items the clusters become more accessible. For example, when a financial company wants to target (future) mortgage owners, probably the best target audience can be found in cluster 3 (persons who are high educated, have a high income, own large houses and always looking for the best mortgages) and cluster 4 (on the eve

of a successful career, with probably a high income, large houses and an increasing demand of mortgages). Or, when there are promotions for families with children, probably the best target audience can be found in clusters 1, 3 and 6;

4. Responsiveness: a cluster must respond uniquely to marketing activities. This criterion is not unequivocally supported in the literature for psychographic segmentation bases (Wedel and Kamakura, 2000, p.14). From Wedel and Kamakura (2000, p.16) it can be concluded that domain specific segmentation bases, like is used in this application, score moderate on the responsiveness criterion. However, the responsiveness of above described cluster solution is supported by the fact that five of the six clusters identified, have successfully been used in differentiated marketing applications. For example, in an application a Dutch energy supplier sent (differentiated) cluster specific questionnaires to all the customers in their database. For each of the five motivational clusters a cluster specific questionnaire was made. Furthermore, a standard (undifferentiated) questionnaire was sent to a control group. In the end it was shown that the response percentages for the groups, that received (differentiated) cluster specific questionnaires, were higher than the response percentage in the control group (Van Hattum and Hoijtink, 2008). In another application about an European mail order company specialized in gardening products, it was shown that, using a randomized experiment, sending cluster specific catalogues to their customers, stimulated buying behavior and increased sales (Van Hattum and Hoijtink, 2009a). See the track record on `www.smartagent.nl` for an overview of projects where these motivational clusters have been used in differentiated marketing activities. From this point of view the six clusters are responsiveness;

5. Stability: a cluster must be stable in time, at least for a period long enough for identification of the clusters, implementation of a differentiated marketing strategy and to produce profitable results. In the literature this criterion is also not unequivocally supported (Wedel and Kamakura, 2000, p.14). However, from a theoretical point of view, the clusters are expected to be stable (Wedel and Kamakura, 2000, p.15). The stability of above described cluster solution is supported by the fact that four of the six clusters identified, have successfully been used in (inter)national marketing and are still workable in day-to-day business (see the track record on `www.smartagent.nl` for an overview of the projects where these motivational clusters have been used). From this point of view, the six clusters are stable, and given the number of projects involved, have been successful and profitable;

6. Actionability: a cluster and the differentiated marketing strategy must be consistent with the goals and core competencies of the company. Because the BSR questionnaire for the domain housing do not only contain psychographic items, but also domain specific items, like, housing preferences, price preferences, branding of new housing projects, et cetera, the six motivational clusters are actionable. The full descriptions of the six clusters provide relevant information on how to communicate with them and to set-up all kind of (differentiated) marketing activities. For example, in order to:

   - determine what kind of houses need to be build and where to locate them (terraced houses with parks and playgrounds for cluster 1, large, luxury houses for cluster 3 and apartments in the city center for cluster 4);
   - to (cross/up) sell mortgage products, that is to stimulate orders in (other and/or more) mortgage products (clusters 3 and 4);
   - determine what kind of facilities, like shops, playgrounds, parks, transportation, schools, are desired and where to locate them (schools and playgrounds in the neighborhood of where persons in cluster 1 live, more parking places for the in general higher average number of vehicles per person in cluster 3, et cetera);
   - do promotions for empty nesters (cluster 5);
   - do promotions for amusement parks or funfairs (clusters 1 and 6).

As Wedel and Kamakura (2000, p.16) conclude, domain specific psychographical segmentations are in general the most effective segmentation studies. Above described BSR segmentation study in the domain housing is such a domain specific psychographical segmentation. Using the six criterion of good market segmentation it is shown that the solution with six motivational clusters is effective and profitable. As such the statistically optimal solution with 35 clusters is reduced to the optimal solution for the intended marketing purposes.

3.5 Application

3.5.1 Dutch Mail Order Company

A Dutch mail order company, specialized in office, workplace and warehouse supplies, is active in the business-to-business market. In this market the mail order company has successfully used market segmentation; customers from the mail order company are divided according to their business activity and their RFM[†]. However, segmentation based on business activity and RFM do not optimally use the available ordering data. In order to optimally use the available ordering data, a new market segmentation study is performed using all the ordering data from the years 2001 to 2005. The goal of this new market segmentation study is to find groups of businesses with (more or less) the same ordering behavior. Within these groups the mail order company wants to set-up (differentiated) marketing activities, like, for example, differentiated promotions, cross or up-selling strategies, or loyalty programs.

The data that are used in this application, come from the mail order company's ordering database. In this database the ordering behavior of 45,610 businesses is recorded for the years 2001 to 2005. In total the mail order company sells more than 50,000 products, classified into 399 product categories. Examples of these product categories are: dust-bins, cups, plates, protective clothing, tape, storage boxes, buckets, hat-racks, fire extinguishers, office desks/chairs/lamps, labellers, staples, tools, cupboards, bookcases, clocks, overhead projectors, cleaning equipments, archive accessories, white boards, black boards, safes, alarm systems, envelopes, telephone/fax machines, et cetera.

After applying the model based clustering algorithm, the clusters found are described using all kind of secondary data. For example: when was the order, how much money was involved, do they get discount, what kind of business is it, do they have a contract, is it a key account, et cetera. What the different groups of businesses are for the Dutch mail order company and how they can be described using the items, is described in the next subsection.

3.5.2 Results

Running the model based clustering algorithm with above described data set and a cluster model with the 399 product categories, renders a globally optimal solution with $Q_{max} = 19$ clusters. Or, in other words, 19 groups of businesses with more or less the same ordering behavior are found. Unfortunately this number of clusters is not desired for the mail order company; it costs too much time and money to interpret all these clusters and to set-up up (differentiated) marketing strategies for these clusters. For the same reasons as in Section 3.4.2, the researcher decides to describe and interpret the reduced solutions with 15 clusters and the solutions with smaller number of clusters. Again, this is a rather subjective decision, but it is an assessment between domain knowledge, available time and money. In order to

[†] RFM stands for: Recency - When was the last order? Frequency - How many order have they placed with us? Monetary value - What is the value of their orders?

determine what $Q_{desired}$ is, the cluster solution with 15 clusters is further reduced until the most effective and most profitable solution is found.

According to main implication of the two conjectures, as described in Section 3.3.2, with a globally optimal solution of $Q_{max} = 19$ latent clusters, all other solutions with $Q < 19$ latent clusters are known to be locally optimal solutions. Using the algorithm, as described in Section 3.4.1, the statistically optimal solution with $Q_{max} = 19$ is reduced to a cluster solution, suited for the intended marketing purposes.

When interpreting the locally optimal cluster solutions, the reduced solution with $Q_{desired} = 10$ clusters is the most effective and most profitable one. This is shown using the six criteria of good market segmentation:

1. Identifiability: a cluster must be clearly defined. It must be clear who is in the cluster. Table 3.5 shows for each product category (for simplification only a few product categories are shown) the probability per cluster. Using these probabilities and the general, observable, descriptive items in Table 3.6 (for simplification only a few items are shown), each of the ten clusters is described. This results in the following short cluster descriptions:

   - Cluster 1: businesses from this cluster are in general more destination buyers; they only visit the mail order company for specific product categories. In this cluster more businesses can be found, that are active in education, care, commercial services and ideal organization. These businesses were occasional buyers; on average 3 orders in 6 product categories were made in 2005. The ordered product categories can be summarized as: durable office furniture, such as desks, chairs, bookcases, et cetera, and office supplies, such as archive maps/cases/boxes. The percentage of businesses that favors fixed discounts was moderate (=14%). In each order the average amount spent was 346 Euro, which is moderate compared to the other clusters. Also a large percentage of the businesses in this cluster did not order a product category in the last year (=49%). Although businesses from this cluster were not big spenders, a relatively large percentage of businesses is called key account (=16%). In 2005 the contribution of business from this cluster to the total sales of the mail order company was 4%. The average contribution to the total sales per business was relatively low (=847 Euro);
   - Cluster 2: businesses in this cluster are active in factory and reparation services. Mainly in the workplace of these businesses. On average these businesses made 3 orders in 7 product categories in 2005 with the lowest amount spent (=260 Euro). These product categories can be summarized

Table 3.5: Cluster Specific Item Probabilities for the $Q_{desired} = 10$ (or Best Local Mode) Solution.

Cluster	1	2	3	4	5	6	7	8	9	10
Cluster size	7.7%	10.5%	53.0%	3.9%	2.3%	2.0%	10.4%	2.8%	2.9%	4.5%
Archive cases	0.02	0.09	0.02	0.23	0.11	0.14	0.15	0.23	0.31	0.43
Ash trays	0.10	0.05	0.02	0.24	0.36	0.27	0.22	0.50	0.21	0.57
Barrows	0.18	0.03	0.03	0.19	0.57	0.18	0.12	0.30	0.12	0.52
Bicycle storage	0.02	0.04	0.02	0.13	0.09	0.10	0.13	0.21	0.09	0.27
Boxes	0.07	0.13	0.03	0.32	0.23	0.33	0.22	0.41	0.50	0.67
Calculators	0.14	0.01	0.00	0.05	0.59	0.10	0.03	0.14	0.06	0.32
Chairs (reception)	0.06	0.05	0.03	0.23	0.22	0.11	0.23	0.49	0.19	0.49
Cleaning devices	0.04	0.08	0.03	0.27	0.16	0.24	0.24	0.41	0.37	0.62
Dust-bins	0.08	0.03	0.04	0.18	0.21	0.11	0.16	0.30	0.09	0.41
Envelopes	0.06	0.01	0.01	0.09	0.19	0.14	0.09	0.29	0.07	0.36
Fire extinguishers	0.01	0.09	0.01	0.31	0.06	0.06	0.08	0.30	0.27	0.58
Floor protection	0.03	0.22	0.04	0.47	0.16	0.12	0.16	0.33	0.36	0.57
Geographical maps	0.16	0.01	0.00	0.06	0.62	0.07	0.03	0.12	0.07	0.33
Letter boards	0.21	0.04	0.03	0.19	0.62	0.22	0.13	0.26	0.17	0.47
Office desks	0.06	0.01	0.00	0.05	0.12	0.60	0.05	0.15	0.13	0.40
Office sets	0.02	0.22	0.02	0.26	0.10	0.15	0.09	0.19	0.63	0.46
Pallet trucks	0.04	0.09	0.03	0.31	0.15	0.12	0.16	0.30	0.18	0.49
Paper shredders	0.15	0.04	0.02	0.22	0.37	0.31	0.18	0.49	0.19	0.66
Pin boards	0.10	0.21	0.08	0.45	0.39	0.23	0.35	0.70	0.46	0.76
Projection screens	0.01	0.05	0.01	0.13	0.04	0.03	0.04	0.09	0.12	0.21
Rail systems	0.01	0.11	0.01	0.40	0.05	0.03	0.04	0.11	0.12	0.39
Rasters	0.01	0.05	0.01	0.16	0.03	0.07	0.07	0.18	0.31	0.40
Security icons	0.03	0.10	0.03	0.34	0.13	0.12	0.17	0.34	0.27	0.54
Small hand tools	0.00	0.04	0.01	0.15	0.04	0.05	0.04	0.14	0.15	0.36
Stamps	0.11	0.01	0.01	0.05	0.25	0.26	0.04	0.11	0.06	0.28
Staples	0.07	0.02	0.01	0.08	0.24	0.10	0.12	0.34	0.08	0.35
Storage racks	0.03	0.07	0.02	0.25	0.12	0.09	0.13	0.38	0.20	0.55
Surprise papers	0.01	0.05	0.01	0.12	0.08	0.04	0.06	0.16	0.16	0.27
Tables (canteen)	0.06	0.13	0.05	0.26	0.19	0.14	0.19	0.38	0.25	0.46
Tables (reception)	0.05	0.17	0.03	0.22	0.15	0.24	0.11	0.28	0.58	0.55
Transportation boxes	0.06	0.04	0.02	0.22	0.23	0.17	0.18	0.47	0.16	0.58
Water hoses	0.01	0.04	0.01	0.19	0.04	0.04	0.03	0.16	0.15	0.38
Wooden racks	0.09	0.04	0.03	0.15	0.22	0.16	0.15	0.35	0.11	0.40
Work benches	0.02	0.06	0.02	0.23	0.08	0.08	0.14	0.33	0.17	0.46

Table 3.6: Cluster Specific Probabilities for the Observable Items, for $Q_{desired} = 10$ (or Best Local Mode) Solution.

Cluster		1	2	3	4	5	6	7	8	9	10
Cluster size	Percentage	7.7%	10.5%	53.0%	3.9%	2.3%	2.0%	10.4%	2.8%	2.9%	4.5%
	Number of businesses	3,528	4,808	24,163	1,777	1,040	922	4,735	1,282	1,313	2,042
Sales	Contribution sales 2001 - 2005	5%	6%	9%	9%	7%	2%	11%	9%	6%	36%
	Av. amount spent per order 2001-2005 (Euro)	412	260	281	299	531	280	321	358	295	366
	Percentage key accounts 2001 - 2005	14%	5%	9%	8%	18%	11%	12%	20%	7%	32%
	Contribution total sales 2005	4%	6%	11%	9%	6%	2%	11%	9%	7%	35%
	Av. contribution sales per business 2005 (Euro)	847	724	414	1,981	2,410	1,257	1,119	2,532	1,996	6,283
	Av. amount spent per order 2005 (Euro)	346	260	273	288	410	269	298	337	284	346
	Av. number of orders 2005	3	3	2	7	6	5	4	8	7	19
	Percentage key accounts 2005	16%	5%	8%	9%	19%	13%	13%	22%	9%	35%
Orders	Percentage without order in 2005	49%	39%	63%	10%	22%	26%	26%	8%	13%	3%
Relationship	Percentage customer since 1999 - 2000	45%	66%	28%	90%	79%	61%	68%	84%	86%	91%
	Percentage customer since 2001 - 2002	27%	20%	27%	8%	16%	23%	21%	13%	10%	7%
	Percentage customer since 2003 - 2004	20%	11%	25%	2%	4%	12%	9%	3%	3%	2%
	Percentage customer since 2005	8%	3%	20%	0%	0%	3%	2%	0%	1%	0%
Marketing	Percentage account manager	4%	4%	3%	7%	10%	4%	6%	11%	7%	25%
	Percentage contracts	15%	6%	10%	10%	20%	12%	14%	22%	10%	36%
	Percentage fixed discounts	14%	6%	9%	9%	18%	12%	13%	21%	9%	34%
	Average discount percentage	12%	11%	12%	11%	12%	12%	12%	11%	12%	11%
	Percentage participant loyalty program	15%	13%	9%	25%	22%	27%	21%	30%	30%	39%
Products	Av. product categories	6	7	2	23	20	21	12	29	23	57

as packaging materials, like: ropes, envelopes, tape, boxes, et cetera. Although these kind of product categories are relatively fast moving goods, the percentage of businesses, that did not bought a product category in 2005 (=39%), was relatively high. Also the average amount spent in each order (=724 Euro), the percentage of key accounts (=5%) and the percentage of businesses that favors fixed discounts (=6%) are relatively low. The contribution to the total sales of the mail order company is 6%, with on average a sales contribution per business of 724 Euro;

- Cluster 3: this cluster contains the majority of the customers (=53%). Businesses from this cluster can be seen as impulse customers, because in general none of the product categories had a higher probability to be bought, the number of orders in 2005 (=2) was the lowest, with on average 2 product categories bought in each order and the average amount involved in this order was relatively low (=273 Euro). A large percentage of the businesses in this cluster (=63%) did not order a product category in the last year. A higher percentage of businesses (=20%) did their first order in 2005. The percentage of businesses from this cluster that favored fixed discounts (=9%) was relatively low. Also the percentage of key accounts in this cluster (=8%) was the lowest. Although the contribution to the total sales (=10.7 percent), was relatively high for this cluster, the average sales contribution per business (=414 Euro) was the lowest;

- Cluster 4: like the businesses in cluster 2, businesses from this cluster also work in the workplace of factories and reparation services. However, businesses from this cluster ordered more frequently and also other product categories. On average 7 orders were made in 23 product categories in 2005. These product categories were not only packaging materials, but also slow moving products needed at the workplace, like: (flash)lights, cases, boxes, dust-bins, security devices, ladders, tools, et cetera. The average amount spent in each order (=288 Euro), the percentage of key accounts (=9%) and the percentage of businesses that favored fixed discounts (=9%) were relatively low. Although the contribution to the total sales (=9%) was moderate for this cluster, the average sales contribution per business (=1,981 Euro) was relatively high;

- Cluster 5: in this cluster more offices are found. Ordering in general more durable office supplies, such as desks, all kind of chairs, white boards, cupboards, bookcases, et cetera. On average 6 orders were made in 20 product categories in 2005. With 410 Euro spent in each order this is the highest compared with the other clusters. Also the percentage of key accounts (=19%) and the percentage of businesses that favors fixed discounts (=18%) were relatively high. From the total sales of the mail order company 7% was contributed from this cluster, which is moderate. Whereas the average sales contribution per business (=2,410 Euro) was relatively high;

- Cluster 6: like the businesses from cluster 1 and 5, businesses from this cluster are also offices. However, businesses from this cluster ordered more frequently than the businesses in cluster 1 and less frequently than the businesses in cluster 5. Also the product categories were different. Where businesses from cluster 1 and 6 were more focused on durable office supplies, businesses from this cluster were more focused on fast moving office supplies, like: paper, pencils, labels, stamps, tape, et cetera. On average 5 orders were made in 21 product categories in 2005. The amount of money spent in each order (=269 Euro) was the lowest of all clusters. A moderate percentage of businesses were called key accounts (=13%). Also the percentage of businesses that favors fixed discounts (=12%) was moderate. Although the contribution to the total sales of the mail order company (=2%) was the lowest in 2005, the average sales contribution per business (=1,257 Euro) was relatively high;

- Cluster 7: like the businesses from cluster 2 and 4, businesses from this cluster have their main activity in the workplace. Compared to the businesses in cluster 2 and 4, businesses from this cluster order more durable product categories, like: white boards, transportation devices, dust-bins, flip-overs, ladders, et cetera. On average 4 orders were made in 12 product categories in 2005, with a relatively low amount of money spent (=298 Euro). Compared to the businesses in cluster 2 and 4, higher percentages of businesses were key account (=13%) and favored fixed discounts (13%). The contribution of these businesses to the total sales of the mail order company (=11%) and the average contribution per business (=1,119 Euro) were moderate compared to the other clusters;

- Cluster 8: businesses from this cluster can be compared to the businesses in cluster 5. Businesses from this cluster are also more offices, ordering even more durable office supplies, like: first aid kits, advertising boards, information boards, projection screens, et cetera. The big difference between the two clusters is, that business from this cluster are more routine buyers. Businesses from this cluster made on average 8 orders in 29 product categories in 2005 with an average amount spent of 337 Euro.

  Furthermore, only 8% of the businesses did not place an order in 2005, whereas it was 22% for the business in cluster 5. Where the percentages of key accounts and businesses favoring fixed discounts were relatively high in cluster 5, these percentages were even higher for the businesses in this cluster; with 22% of the businesses as key accounts and 21% of the businesses favoring fixed discounts these percentages were among the highest. From the total sales of the mail order company 9% was contributed from businesses from this cluster. The average contribution per business was with 2532 Euro among the highest;

  - Cluster 9: businesses from this cluster look very similar to the businesses in cluster 4. The average number of orders in product categories in 2005 (=7 orders in 23 product categories) were similar. Also the average amount spent in each order (=284 Euro), the percentage of key accounts (=9%), the percentage businesses favoring fixed discounts (=9%), the contribution to the total sales of the mail order company (=7%) and the average contribution per business (=1,996 Euro) were (more or less) similar. The big difference was the product categories ordered. Businesses in this cluster ordered more product categories, like: work clothing, protective gear, fire extinguishers, fire showers, first aid kits, chemicals, glues, kits, et cetera;

  - Cluster 10: for businesses from this cluster the mail order company is the primary place to shop frequently. On average 19 orders were made in 57 product categories in 2005, with an average amount spent of 346 Euro. Both offices and workplaces can be found in this clusters, ordering both durable and fast moving products for offices and workplaces. With 35% of the businesses as key accounts this was the highest compared to other clusters. Also the percentage of businesses favoring fixed discounts (=34%), the contribution to the total sales of the mail order company (=35%) and the contribution per business (=6,283 Euro) were also the highest.

  Given these descriptions the ten clusters are clearly defined. It is clear what kind of business is in the cluster;

2. Substantiality: a cluster must be large enough to ensure the profitability of developing a differentiated marketing strategy. From the row 'Cluster size' in Table 3.5 it can be seen that all ten clusters have substantial weights, that is cluster 1 = 7.7%, cluster 2 = 10.5%, cluster 3 = 53.0%, cluster 4 = 3.9%, cluster 5 = 2.3%, cluster 6 = 2.0%, cluster 7 = 10.4%, cluster 8 = 2.8%,

cluster 9 = 2.9% and cluster 10 = 4.5%. Whether these substantial clusters are profitable enough, can be decided after a differentiated marketing strategy has taken place (Stanton and Pires, 1999). However, the descriptions of the ten clusters provide relevant information on how to communicate with them and to set-up all kind of (differentiated) marketing strategies;

3. Accessability: a cluster must be reachable through promotional or distributional marketing activities. According to the literature this criterion appears to be somewhat limited for the segmentation basis, that is used in this market segmentation study (Wedel and Kamakura, 2000, p.11). However, in above described segmentation study also observable items, such as business activity, number of employees, et cetera, are used in order to further describe the ten clusters. Using these observable items the clusters become more accessible (Frank, 1972). For example, if the target businesses of a differentiated marketing action are offices, probably the best target audience can be found in clusters 1, 5, 6 and 8. Or, if the target businesses are workplaces of companies, probably the best target audience can be found in clusters 2, 4, 7 and 9;

4. Responsiveness: a cluster must respond uniquely to marketing activities. According to the literature, segmentation studies with the segmentation basis, that is used in this application, are to a lesser extent responsive (Wedel and Kamakura, 2000, p.11). However, from Wedel and Kamakura (2000, p.16) it can be concluded that this segmentation basis scores well on the responsiveness criterion. Although no marketing activities for these clusters have been initiated yet, it is supposed that the clusters respond uniquely to marketing activities. The descriptions of the ten clusters provide relevant information on how to communicate with the businesses within the clusters. In the marketing activities you can target either offices or workplaces. Or, offices, that frequently buy either durable supplies or fast moving suppliers. Or, loyal businesses or new businesses. All different groups of businesses that can be targeted using the ten identified clusters, that are supposed to respond uniquely to (differentiated) marketing activities;

5. Stability: a cluster must be stable in time, at least for a period long enough for identification of the clusters, implementation of a differentiated marketing strategy and to produce profitable results. According to the literature segmentation studies with a segmentation basis, that is used in this application, are to a lesser extent stable (Wedel and Kamakura, 2000, p.11). However, from Wedel and Kamakura (2000, p.16) it can be concluded that this segmentation basis scores well on the stability criterion. This is supported by the

fact that the ordering behavior in each cluster is more or less similar over the five years;

6. Actionability: a cluster and the differentiated marketing strategy must be consistent with the goals and core competencies of the company. Because the ten cluster are specifically made for the mail order company, in cooperation with the marketeer, the full descriptions of the ten clusters provide relevant information on how to communicate with them and to set-up all kind of (differentiated) marketing activities. For example, in order to:

    - revitalize businesses that did not order from one of the product categories in the last year (cluster 1 and 3);
    - do specific promotions for new customers (cluster 3);
    - to cross/up sell product categories, that is to stimulate orders in other and/or more product categories (for example, stimulating businesses in cluster 1 to order also fast moving office supplies, like the businesses in cluster 6);
    - to maintain good, profitable businesses (cluster 8 and 10). The prevention against churn is the most important goal for these type of marketing activities.

Using the six criterion of good market segmentation, it is shown that the solution with ten clusters of businesses is effective and profitable. As such the statistically optimal solution with 19 clusters is reduced to the optimal solution for the intended marketing purposes.

3.6 Discussion

In marketing there is a wish of dividing a market as effectively as possible. Resulting in a market segmentation with as few clusters as possible to fully describe the market, and finally, to solve the marketing problem at hand. From this point of view it is clear that it is possible that there may be a difference between the statistically optimal solution and the solution suited for the intended marketing purposes. From a marketeer's point of view this globally optimal solution may contain too many clusters to interpret or to make differentiated strategies for.

This chapter proposed an algorithm to reduce the statistically optimal number of clusters to a smaller number, suited for marketing purposes. Point of departure for this is the statistically optimal solution. Using an agglomerative method and an information criterion, a decision tree is used to combine clusters till a useful number of clusters for the intended marketing purposes is reached.

This chapter illustrated the reduction algorithm using two marketing applications. In the first application the Dutch housing market was segmented. Using a model based clustering algorithm it turned out that the statistically optimal cluster solution was a solution with 35 clusters. Or, in other words, the total housing market in The Netherlands could be divided into 35 clusters; 35 different groups of customers, that had the same views, motivations and attitude with respect to housing. But as time and money was limited, interpreting, describing or making differentiated marketing plans for these 35 clusters, was too expensive. Using the domain knowledge of a real estate consultant, there was a wish to reduce the number of clusters to a smaller number suited for marketing purposes. Using the proposed reduction algorithm the statistically optimal solution with 35 clusters was reduced to a solution with six clusters. Using six criteria of good market segmentation this reduced cluster solution was evaluated and was found to be profitable and effective for further (differentiated) marketing activities.

In the second application the ordering database of a Dutch mail order company was segmented. Using a model based clustering algorithm it turned out that the statistically optimal cluster solution was a solution with 19 clusters. Or, in other words, 19 groups of businesses with more or less the same ordering behavior were found. However, for the same reasons as in the first application, this was not desired. Using the domain knowledge of the mail order company's marketeer, there was a wish to reduce the number of clusters to a smaller number, suited for the intended marketing purposes. Using the proposed reduction algorithm, the statistically optimal solution with 19 clusters was reduced to a solution with ten clusters. Using six criteria of good market segmentation, these ten cluster were found to be effective and profitable.

This chapter showed that, using six criteria of good market segmentation, an information criterion and two conjectures, describing the geometry of model based cluster models, the statistically optimal solution can be reduced to a solution suited for the intended marketing purposes.

Appendix I

The BSR Questionnaire
When you think of a particular car you may also think of the typical person driving such a make or model. Now, when you think of housing* and of the way in which you live (or would like to live), what kind of people go with that? How do these people live, what are their character traits, their hobbies, occupations, et cetera?

Q1. Which character traits fit the best for the person that has the same opinion about housing as you do?

○ a little bit shy
○ adventurous
○ capable
○ cozy
○ energetic
○ a little bit imprudent
○ honest
○ jovial
○ opinionated
○ self-assured
○ serious
○ spontaneous

○ a little impatient
○ assertive
○ cheerful
○ critical
○ enthusiastic
○ gentle
○ intelligent
○ sympathetic
○ ordinary
○ self-confident
○ down-to-earth
○ strong character

○ easygoing
○ balanced
○ classy
○ deliberate
○ leader
○ helpful
○ interested in others
○ neat
○ passionate
○ serene
○ commercial

Q2. Which family or household types fit the best for the person that has the same opinion about housing as you do?

○ a family where everyone goes their own way
○ bachelor
○ busy dynamical family
○ happy family
○ ideal family
○ not suited for family life
○ quiet family
○ single
○ stable family
○ striving for a family

○ artistic household
○ broad-minded family
○ cozy old-fashioned family
○ harmonious family
○ isolated family
○ perfect family
○ rigid family
○ sportive family
○ aristocratic household
○ warm family

* ”housing” may be replaced by the subject of the study, for example, energy, financial services, insurance, health care, et cetera.

Q3. Which occupations fit the best for the person that has the same opinion about housing as you do? Occupations can be done both by males or females.

○ account manager
○ member of the board
○ commercial assistant
○ e-business
○ free-lancer
○ journalist
○ no occupation
○ photographer
○ programmer
○ receptionist
○ shop assistant
○ sports teacher
○ temporary employee
○ vets assistant
○ activity guide
○ business-man/-woman
○ commissioner
○ entrepreneur
○ full time house wife
○ male nurse
○ nurse
○ artist
○ project manager
○ scientist
○ shopkeeper
○ student
○ truck driver
○ volunteer
○ beauty specialist
○ social worker
○ designer
○ financial planner
○ house-husband
○ manager
○ part time house-wife
○ anchor man
○ public servant
○ secretary
○ social worker
○ stylist
○ unemployed

Q4. Which hobbies, interests and/or leisure activities fit the best for the person that has the same opinion about housing as you do?

○ a sociable evening with friends
○ top-notch achievement
○ build a successful career
○ classy parties
○ do odd jobs around the house
○ going to a discotheque
○ make dreams come through!
○ playing chess
○ snow boarding
○ visiting friends and relatives
○ watching TV
○ active sports
○ astrology
○ camping
○ a day out
○ gardening
○ golf
○ religious matters
○ reading magazines
○ working out
○ team sports
○ adventurous holidays
○ being at home quietly
○ cars / motor bikes
○ dine out together
○ going out together
○ investing in stocks
○ swimming
○ shopping
○ surfing the Internet
○ visiting a pub

Q5. Which values fit the best for the person that has the same opinion about housing as you do?

○ anonymity
○ enthusiasm
○ heroism, glory
○ passion
○ recognition of performances
○ self-belief
○ social harmony
○ success in life
○ challenge, stimulation
○ expression, uniqueness
○ independence
○ privacy, tranquility
○ respect
○ self-expression, growth
○ solidarity
○ enjoyable life
○ friendship
○ intimacy
○ rationalism
○ security
○ social alliance
○ status

Chapter 4

A Comparison of Model Based Clustering Algorithms

Abstract

Most of the papers about model based clustering mention the advantages of the probabilistic approach over standard clustering. Few papers can be found that compare model based clustering approaches with standard clustering. Papers that actually compare different model based clustering approaches are even scarcer. This chapter compares three different model based clustering approaches: a Bayesian model based clustering approach and the approaches implemented in LatentGold and Glimmix. Using simulation studies the performance of each of the approaches is evaluated.

4.1 Introduction

In recent years model based clustering has become a popular technique, resulting in numerous papers with specific model based clustering approaches and their applications (Fraley and Raftery, 1998; Hoijtink and Notenboom, 2004; Ter Braak et al., 2003; Van Hattum and Hoijtink, 2009b; Vermunt and Magidson, 2000; Wedel and Kamakura, 2000, p.97). Most of these papers briefly mention the advantages of a model clustering approach over standard clustering (Hair et al., 1984, p. 469-518). There are few papers that compare model based clustering with standard clustering (DiStefano and Kamphaus, 2006; Magidson and Vermunt, 2002; Wang et al., 2008). Papers that actually compare model based clustering approaches are even scarcer (Meila and Heckerman, 2001; Ter Braak et al., 2003).

This chapter compares three model based clustering approaches: a Bayesian model based clustering approach (Hoijtink and Notenboom, 2004; Van Hattum and Hoijtink, 2009b) and the approaches implemented in the software packages LatentGold (Vermunt and Magidson, 2000) and Glimmix (Wedel and Kamakura, 2000, p.181-186). In order to evaluate the three clustering approaches, different data sets with both locally independent items and locally dependent items within the latent clusters are simulated. The performances of the clustering approaches are evaluated using diagnostics that will be specified in the sequel.

The structure of this chapter is as follows. Section 4.2 describes the model based clustering algorithms implemented in LatentGold and Glimmix. Furthermore, a third algorithm, that is Bayesian model based clustering, is described. Section 4.3 compares and evaluates these algorithms using several simulated data sets. This chapter concludes with a discussion in Section 4.4.

4.2 Model Based Clustering Approaches

4.2.1 Introduction

An important difference between standard clustering (Hair et al., 1984, p. 469-518) and model based clustering (Banfield and Raftery, 1993; Bensmail et al., 1997; Fraley and Raftery, 1998; Newcomb, 1886; Pearson, 1894; Vermunt and Magidson, 2000, p. 1-2, 152) is that in the latter it is assumed that the data are generated by a certain mixture of underlying probability distributions.

An advantage of this probabilistic approach is that the cluster criterion (Hair et al., 1984, p. 482-490; Wedel and Kamakura, 2000, p. 39-73), which is usually difficult to define and calculate for complex models, is not needed. A further advantage of this approach is that uncertainty about a respondent's cluster membership is taken into account. A disadvantage of the model based clustering approach is that, although today's computers have extremely fast processors, it may take a while to run a model based clustering algorithm.

This section briefly describes three different model based clustering approaches. Each with their own advantages and disadvantages.

4.2.2 Likelihood

Let x_{ij} denotes the response of respondent $i = 1, \ldots, N$ to item $j = 1, \ldots, J$, $x_{ij} \in \{0, 1\}$, where 1 indicates that respondent i picked item j and 0 indicates that respondent i did not pick item j. The $N \times J$ matrix $\boldsymbol{X}$ contains the item responses. The J vector $\boldsymbol{x}_i$ is defined as a vector containing the response pattern or item responses of respondent i. The N vector $\boldsymbol{x}_j$ is defined as a vector containing the responses of the respondents to item j. The data matrix $\boldsymbol{X}$ is split into a data matrix $\boldsymbol{X}^0$ and $\boldsymbol{X}^*$, that is $\boldsymbol{X} = \{\boldsymbol{X}^0, \boldsymbol{X}^*\}$ and $J = J^0 + J^*$. The part $\boldsymbol{X}^0$ is the $N \times J^0$ data matrix containing J^0 items that, within the latent clusters, are independent of the other items. The part $\boldsymbol{X}^*$ is the $N \times J^*$ data matrix containing the J^* items that, within the latent clusters, are dependent on some of the other items. Note, that $\boldsymbol{x}_i = \{\boldsymbol{x}_i^0, \boldsymbol{x}_i^*\}$.

Within each cluster $q = 1, \ldots, Q$ each of the J^0 locally independent items is characterized by a parameter $\pi_{j|q}$, that is the probability of responding 1 to item j in cluster q. Note that $\boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_q, \ldots, \boldsymbol{\pi}_Q\}$ and $\boldsymbol{\pi}_q = \{\pi_{1|q}, \ldots, \pi_{j|q}, \ldots, \pi_{J^0|q}\}$.

Within each cluster $q = 1, \ldots, Q$ a log-linear model, containing main effects and two-way interaction effects, is used to model the responses in $\boldsymbol{X}^*$. The parameters of these log-linear models are denoted by $\boldsymbol{\lambda}_q$. A further elaboration of these log-linear models follows in the sequel.

Let $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_q, \ldots, \omega_Q\}$ be the Q vector containing the cluster weights, that is the proportion of persons allocated to each cluster and let $\omega_{q|i}$ denotes the probability that respondent i belongs to latent cluster q. The N vector $\boldsymbol{\tau}$ contains the unobserved cluster memberships for each person $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_i, \ldots, \tau_N\}$, where $\tau_i \in \{1, \ldots, Q\}$. Finally, the matrix $\boldsymbol{M}$ is a $N \times J$ indicator matrix with elements m_{ij}, where a 0 indicates that a response is missing and a 1 that a response is observed.

The general form of the data likelihood of the model based cluster model is given by

$$L(\boldsymbol{X} \mid \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\omega}) = \prod_{i=1}^{N} \sum_{q=1}^{Q} \omega_q P(\boldsymbol{x}_i \mid \tau_i = q), \tag{4.1}$$

where

$$P(\boldsymbol{x}_i \mid \tau_i = q) = P(\boldsymbol{x}_i^0 \mid \tau_i = q) P(\boldsymbol{x}_i^* \mid \tau_i = q). \qquad (4.2)$$

For the locally independent items

$$P(\boldsymbol{x}_i^0 \mid \tau_i = q) = \prod_{j=1}^{J^0} P(x_{ij}^0 \mid \tau_i = q), \qquad (4.3)$$

with

$$P(x_{ij}^0 \mid \tau_i = q) = \pi_{j|q}^{x_{ij}^0} (1 - \pi_{j|q})^{1 - x_{ij}^0}. \qquad (4.4)$$

For the J^* dichotomeous items that have local dependencies within the clusters the number of possible response vectors is 2^{J^*} and is denoted by $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p, \ldots, \boldsymbol{y}_{2^{J^*}}\}$ and $\boldsymbol{y}_p = \{y_{p1}, \ldots, y_{pj}, \ldots, y_{pJ^*}\}$. Let the number of elements of $\boldsymbol{\lambda}_q$ be denoted by L. Let $\boldsymbol{R}$ denotes a $2^{J^*} \times L$ design matrix. Then

$$P(\boldsymbol{x}_i^* \mid \tau_i = q) = \frac{exp^{\boldsymbol{R}_p \boldsymbol{\lambda}_q}}{\sum_{p'=1}^{2^{J^*}} exp^{\boldsymbol{R}_{p'} \boldsymbol{\lambda}_q}}, \qquad (4.5)$$

where $\boldsymbol{R}_p$ denotes the row from $\boldsymbol{R}$ for which $\boldsymbol{x}_i^* = \boldsymbol{y}_p$. The interested reader is referred to, for example, Schafer (1997, p. 289-292) for a further and more general elaboration. For illustrative purposes we restrict ourselves to a simple elaboration with a set $\boldsymbol{X}^*$ with $J^* = 3$ and $\boldsymbol{\lambda}_q = \{\lambda_{0,q}, \lambda_{1,q}, \lambda_{2,q}, \lambda_{3,q}, \lambda_{1,2,q}, \lambda_{1,3,q}\}$, that is $L = 6$. Note that within each cluster this model contains an intercept, three main effects and two two-way interactions. Here

$$\boldsymbol{Y} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

and

$$\boldsymbol{R} = \begin{pmatrix} 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

4.2.3 Geometry of Cluster Likelihoods

An important aspect in model based clustering is the protection against local maxima. In general clustering algorithms can converge on a local mode rather than converging on the globally best solution. Two types of local maxima can be encountered. In the first type the number of clusters in the cluster solution is smaller than the true number of clusters. The second type arises when the number of clusters in the cluster solution is too small. As will elaborated in the sequel the likelihood then has multiple maxima and only one of them is the 'best' local maximum for the number of clusters at hand. A global maximum solution is a solution with the right number of clusters and has accurate estimates of the parameters.

To illustrate the existence of local modes of the second type, a data set containing 1000 respondents and 10 locally independent items is simulated. In the simulated data set the global maximum $Q_{max} = 3$. See Table 4.1 for the parameters used to simulate the data set, that are cluster weights and cluster specific probabilities.

According to two conjectures of Hoijtink and Notenboom (2004), with $Q_{max} = 3$ the number of possible mixtures is five (one mixture of three clusters, three mixtures of two clusters in which two of the three clusters are combined and one mixture of one cluster). To illustrate that there are three different local maxima with $Q = 2$ clusters, the clustering algorithm* is run multiple times with different random initializations. See Table 4.2 for the parameter estimates resulting from each of the three random initializations. In the first random initialization Cluster 2 and 3 from Table 4.1 are combined. Or, in other words, the probabilities of Cluster 2 from Table 4.2 are the weighted (using cluster weights) averages of the probabilities of Cluster 2 and 3 from Table 4.1. The cluster specific probabilities for Cluster 1 in the

*In this example the Bayesian approach, which is described in Section 4.2.6, is used. However, the same results can be obtained by using, for example, LatentGold (described in Section 4.2.4) with multiple runs with random initializations.

Table 4.1: Parameters Used to Simulate a Data Set to Illustrate Local Modes.

	Population		
Cluster q	1	2	3
ω_q	0.330	0.330	0.340
$\pi_{1\|q}$	0.80	0.10	0.10
$\pi_{2\|q}$	0.80	0.10	0.10
$\pi_{3\|q}$	0.80	0.10	0.10
$\pi_{4\|q}$	0.10	0.80	0.10
$\pi_{5\|q}$	0.10	0.80	0.10
$\pi_{6\|q}$	0.10	0.80	0.10
$\pi_{7\|q}$	0.10	0.10	0.80
$\pi_{8\|q}$	0.10	0.10	0.80
$\pi_{9\|q}$	0.10	0.10	0.80
$\pi_{10\|q}$	0.50	0.50	0.50

first column of Table 4.1 are more or less equal to the cluster specific probabilities for Cluster 1 in the first column of Table 4.2. Likewise it can be seen that in the second random initialization Cluster 1 and 3 are combined and in the third random initialization Cluster 1 and 2 are combined.

This illustrates that there are three different local maxima with $Q = 2$ clusters. Precautions to avoid local maxima are discussed later in this section.

4.2.4 LatentGold

Introduction

A well-known software package for model based clustering is the program LatentGold[†] by Vermunt and Magidson (2000). This package is the Windows version of the program LEM by Vermunt (1997).

An important difference between LatentGold and LEM is that LatentGold has faster (EM followed by Newton-Raphson) and safer (sets of starting values) estimation methods for model based clustering.

Other nice features of LatentGold are the allowance for items with different scale types, such as nominal, ordinal, continuous, or any mixture of these types. In

[†] In this chapter LatentGold version 4.5 is used

Table 4.2: Parameter Estimates Resulting from Three Different Random Initializations.

	Initialization 1		Initialization 2		Initialization 3	
Cluster q	1	2	1	2	1	2
ω_q	0.317	0.683	0.325	0.675	0.331	0.669
$\pi_{1\|q}$	0.85	0.09	0.10	0.44	0.08	0.46
$\pi_{2\|q}$	0.81	0.10	0.09	0.45	0.10	0.44
$\pi_{3\|q}$	0.84	0.09	0.08	0.45	0.08	0.45
$\pi_{4\|q}$	0.09	0.45	0.82	0.10	0.10	0.45
$\pi_{5\|q}$	0.11	0.46	0.81	0.13	0.15	0.45
$\pi_{6\|q}$	0.10	0.44	0.79	0.11	0.14	0.43
$\pi_{7\|q}$	0.10	0.45	0.13	0.44	0.81	0.11
$\pi_{8\|q}$	0.11	0.44	0.13	0.44	0.79	0.12
$\pi_{9\|q}$	0.10	0.45	0.12	0.45	0.83	0.10
$\pi_{10\|q}$	0.52	0.55	0.56	0.53	0.54	0.54

this chapter only nominal items with two answer categories are used, the so-called dichotomous items.

LatentGold produces informative output in a very structured way, such as, cluster weights, cluster specific probabilities, bi-variate residuals, information criteria and iteration details.

Likelihood and Parameter Estimation

The likelihood of the data in LatentGold is given in (4.1). In order to estimate the cluster specific probabilities ($\boldsymbol{\pi}$), cluster specific log-linear parameters ($\boldsymbol{\lambda}$) and cluster weights ($\boldsymbol{\omega}$), LatentGold makes use of Maximum Likelihood (ML). To find the ML estimates, the program uses two well-known algorithms: EM (Dempster et al., 1977) and Newton-Raphson (Haberman, 1988).

The EM algorithm is an iterative algorithm that contains, for the models that have locally independent items within the latent clusters, the following steps:

Initialization

1. In the very first iteration of the EM-algorithm the respondents are randomly divided into Q clusters.

E-step

2. $\omega_{q|i} = \frac{\omega_q P(\boldsymbol{x}_i^0|\tau_i=q)}{\sum_{q'=1}^{Q}\omega_{q'}P(\boldsymbol{x}_i^0|\tau_i=q')}$, for $q = 1,\ldots,Q$ and $i = 1,\ldots,N$.

M-step

3. $N_q = \sum_{i=1}^{N}\omega_{q|i}$, for $q = 1,\ldots,Q$.
4. $\omega_q = \frac{N_q}{N}$, for $q = 1,\ldots,Q$.
5. When the items are locally independent: $\pi_{j|q} = \frac{\sum_{i=1}^{N}\omega_{q|i}m_{ij}x_{ij}^0}{\sum_{i=1}^{N}m_{ij}x_{ij}^0}$, for $j = 1,\ldots,J^0$ and $q = 1,\ldots,Q$.

For the models that have locally dependent items within the latent clusters, the EM-algorithm contains the following steps:

Initialization

1. In the very first iteration of the EM-algorithm the respondents are randomly divided into Q clusters

E-step

2. $\omega_{q|i} = \frac{\omega_q P(\boldsymbol{x}_i^*|\tau_i=q)}{\sum_{q'=1}^{Q}\omega_{q'}P(\boldsymbol{x}_i^*|\tau_i=q')}$, for $q = 1,\ldots,Q$ and $i = 1,\ldots,N$.

M-step

3. $N_q = \sum_{i=1}^{N}\omega_{q|i}$, for $q = 1,\ldots,Q$.
4. $\omega_q = \frac{N_q}{N}$, for $q = 1,\ldots,Q$.
5. When the items j and k are locally dependent: $\pi_{j,k|q}^{a,b} = \frac{\sum_{i=1}^{N}\omega_{q|i}I_{j=a,k=b}}{N_q}$, for $j = 1,\ldots,J^*, k = j+1,\ldots,J^*$ and $q = 1,\ldots,Q$. Here, $\pi_{j,k|q}^{a,b}$ is the probability of responding a to item j and responding b to item k in cluster q. Note that $a, b \in \{0,1\}$. The indicator function $I_{j=a,k=b}$ is 1 if $x_{ij} = a$ and $x_{ik} = b$ and 0 otherwise. Using iterative proportional fitting (Schafer, 1997, p.298-299) these probabilities $\pi_{j,k|q}^{a,b}$ are transformed to log-linear parameters.

Note that for models, that contain both locally independent and locally dependent items within the latent clusters, the last EM algorithm can be used.

A problem with the EM algorithm is when to stop it. The EM algorithm stops when the likelihood or, in the case of LatentGold, the parameters hardly change from one iteration to the next. However, Wedel and Kamakura (2000, p. 88) describe that this is a lack of progress, rather than a measure of convergence and that there is evidence that the EM-algorithm is often stopped too early. In order to avoid this problem, LatentGold uses the speed of Newton-Raphson when close to the optimal solution.

Size of Data Sets

In LatentGold there is no limit concerning the number of respondents and the number of items in a data set. As is shown in Section 4.3 data sets with 50.000 respondents and more than 150 items can be analyzed. Furthermore, according to the web site of LatentGold (`www.statisticalinnovations.com`) LatentGold runs 20 or more times faster than other model based clustering programs.

Number of Clusters

The problem of identifying the number of latent clusters is still without a satisfactory statistical solution and one of the main research topics in model based clustering (Wedel and Kamakura, 2000, p. 91).

The most popular method of determining the number of latent clusters is by using the information criteria BIC and CAIC (Wedel and Kamakura, 2000, p. 91). Both information criteria are among the standard output from LatentGold and can be used to determine the number of latent clusters. The researcher pre-specifies a range of cluster solutions. The cluster solution with the lowest value of the information criterion is preferred, because information criteria can be seen as the distance between the current model and the true model.

Local Maxima

In order to prevent against local maxima, LatentGold initializes with multiple random starting values. In other words, the initialization step in the EM-algorithm is done multiple times. LatentGold offers the possibility to specify a number of random starting values. Although there is no guarantee that the optimal solution is found, this procedure increases the probability of finding the optimal solution (Vermunt and Magidson, 2000, p. 168). Note that you may find a type 2 local maximum, as described in Section 4.2.3. Research by Ter Braak et al., (2003) has shown that repeated runs with LatentGold on a single data set can lead to different conclusions in terms of number of clusters, even with an increase in number of random initial configurations.

The problem of local maxima can also occur because LatentGold tries to find an optimal solution within a pre-specified range of number of clusters (e.g. give me the optimal solutions with two to five clusters). It may be clear that, although the user may find the best solution within this range, the global maximum (especially for large data sets) may be outside this range. Note that this is a type 1 local maximum, as described in Section 4.2.3.

Missing Values

LatentGold has two options to deal with missing values. The first option is to exclude respondents with missing values from the analysis, the so-called available case analysis. As a result the respondents with missing values are not used for the parameter estimation and no cluster memberships are assigned to them.

The second option is that respondents with missing values are included in the analysis (under the assumption that the missingness is missing at random (MAR) (Schafer and Graham, 2002)). Only the available information is taken into account when estimating the model parameters. The EM-algorithm shows how the missing indicator m_{ij} is used to deal with missing values in models assuming local item independence within the latent clusters. From personal communication with the developer of LatentGold we know that LatentGold also deals with missing values in models assuming local item dependencies within the latent clusters. However, the algorithm used has not been published yet.

Local Item Dependence

The basic assumption in model based clustering is that the items are locally independent within the latent clusters. This assumption is often violated (Ainslie and Rossi, 1998; Manchanda et al., 1999; Reboussin et al., 2008; Van Hattum and Hoijtink, 2009b). A practical way to handle this is to increase the number of clusters until a model with an acceptable fit is obtained (Vermunt and Magidson, 2000, p.155). An alternative is to relax this assumption (Ainslie and Rossi, 1998; Hagenaars, 1988; Manchanda et al., 1999; Vermunt and Magidson, 2000, p.155; Van Hattum and Hoijtink, 2009b).

LatentGold has the possibility to incorporate locally dependent items. The probability $P(\boldsymbol{x}_i^* \mid \tau_i = q)$ in (4.2) is defined as in (4.5), where the cluster specific vectors $\boldsymbol{\lambda}_q = \{\lambda_{0,q}, \lambda_{1,q}, \ldots, \lambda_{j,q}, \ldots, \lambda_{J^*,q}, \lambda_{\cdot,\cdot,q}, \ldots, \lambda_{\cdot,\cdot,q}\}$, for $q = 1, \ldots, Q$. Note that within each latent cluster the model contains an intercept, J^* main effects and (a subset of) the two-way interaction effects.

Using Step 5 in the EM-algorithm for models with items that are locally dependent, the cluster specific log-linear parameters $\boldsymbol{\lambda}_q$ are estimated.

One of the nice features in LatentGold is the possibility of detecting locally dependent items by means of bi-variate residuals. These bi-variate residuals, sometimes referred to as modification indices, indicate how similar the estimated and observed bi-variate associations are (Vermunt and Magidson, 2000, p.174). For large residuals the model under study can be accounted for by adding the corresponding two-way interaction in the model.

4.2.5 Glimmix

Introduction

Another software package for model based clustering is the program Glimmix by Wedel (1997). Glimmix has resemblance with LatentGold, both in estimation and output. Although Glimmix can handle different link functions and distributions, this chapter deals with dichotomous data, thus we only discuss the binomial distribution with a logit link function.

Likelihood and Parameter Estimation

Because Glimmix only deals with models with local item independence, the likelihood of the data that is used in Glimmix is given in (4.1), with accompanying probabilities in (4.3) and (4.4).

Parameter estimation in Glimmix is done using the EM algorithm, as described in Section 4.2.4. Contrary to LatentGold, Glimmix only uses the EM algorithm, whereas LatentGold uses EM and Newton-Raphson consecutively.

Size of Data Sets

According to Wedel (1997, p.6) Glimmix is able to analyze up to 150 locally independent items and up to 50.000 respondents. Furthermore, Glimmix has a special option for data sets containing a large number of respondents. As is shown in Section 4.3, computation time is increased considerably when analyzing large data sets. Wedel (1997, p.29) states that not much is lost by analyzing a sample from the large data set. In other words, Wedel (1997, p.29) states that if the clustering model is run on a smaller sample, randomly drawn from the large data set, convergence is considerably faster and not much is lost in a statistical sense.

The excluded respondents that are not in the clustering model are automatically assigned to the latent clusters by calculating the $\omega_{q|i}$'s for these respondents.

In the case that the researcher wants to analyze a data set that exceeds the maximum of 50.000 respondents, Glimmix computes the fraction of the respondents that can still be analyzed and randomly samples that fraction from the data. In order to speed up computation time for these cases, Glimmix uses an EM algorithm with a somewhat weaker convergence criterion.

Number of Clusters

As mentioned in Section 4.2.4 the number of latent clusters has to be inferred from the data. Like LatentGold, Glimmix calculates different information criteria, among others, the BIC and CAIC. Wedel (1997, p.32) prefers to use the CAIC or BIC.

The way to find the number of clusters is the same as for LatentGold (Section 4.2.4). The researcher pre-specifies a range of cluster solutions. The solution with the lowest value for the information criterion is preferred. To make it easier for the researcher, Glimmix provides a plot of the information criteria against the number of latent clusters to allow for a quick assessment of the appropriate number of latent clusters (Wedel, 1997, p.31-32).

Local Maxima

The way to deal with local maxima in Glimmix is the same as in LatentGold.

Missing Values

Glimmix has two options to deal with missing values. The first option is available case analysis, in which respondents with missing values are excluded from the analysis. No cluster memberships are assigned to these respondents.

The second option is to replace the missing values by a certain fixed value. In the case of this chapter the missing values will be replaced by the response on the item at hand with the highest frequency. Note that this is a form of mean substitution (Schafer and Graham, 2002) for categorical items.

Local Item Dependence

As far as we know, the modelling of items, that are locally dependent within the latent clusters, is not possible in Glimmix.

4.2.6 Bayesian Mixtures of Log-Linear Models

Introduction

The previous sections show that the LatentGold and Glimmix use, among others, the EM-algorithm in order to estimate the model parameters. In this section the parameters are estimated using a Markov Chain Monte Carlo (MCMC) method. The MCMC method renders a sample from the global mode of this posterior distribution. From this sample the expected a posteriori (EAP) estimates (Hoijtink, 2000) for each parameter can easily be calculated.

Likelihood and Parameter Estimation

The posterior distribution of the parameters of the cluster model is proportional to the product of the likelihood and the prior distribution. The likelihood of the data is given in (4.1). The prior distribution is based on standard uninformative and mutually independent Dirichlet distributions for the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\omega}$. Stated otherwise, this prior has a density that is constant and independent of the values $\boldsymbol{\pi}$ and $\boldsymbol{\omega}$.

As is elaborated in the sampling algorithm below, the log-linear parameters $\boldsymbol{\lambda}_q$ are sampled using the probabilities $P(\boldsymbol{y}_p \mid \tau_i = q)$, for $p = 1, \ldots, 2^{J^*}$. For ease of notation we denote the relation between $\boldsymbol{\lambda}_q$ and $P(\boldsymbol{y}_p \mid \tau_i = q)$ by $\boldsymbol{\lambda}(\boldsymbol{.})_q$. Note that in the sequel the probabilities $P(\boldsymbol{y}_p \mid \tau_i = q)$ are sampled and that a standard uninformative Dirichlet distribution is used (Schafer, 1997, p. 306) as the prior distribution for $P(\boldsymbol{y}_p \mid \tau_i = q)$. The prior distribution for the cluster model becomes:

$$h(\boldsymbol{\pi}, P(\boldsymbol{y}_p \mid \boldsymbol{\tau} = 1), \ldots, P(\boldsymbol{y}_p \mid \boldsymbol{\tau} = Q), \boldsymbol{\omega}) \propto \textit{constant}. \tag{4.6}$$

The prior $h(\boldsymbol{\pi}, \boldsymbol{\lambda}(\boldsymbol{.})_1, \ldots, \boldsymbol{\lambda}(\boldsymbol{.})_Q, \boldsymbol{\omega})$ follows from (4.6).

As a result the posterior distribution of the parameters of the cluster model is

$$\textit{Post}(\boldsymbol{\pi}, \boldsymbol{\lambda}(\boldsymbol{.})_1, \ldots, \boldsymbol{\lambda}(\boldsymbol{.})_Q, \boldsymbol{\omega} \mid \boldsymbol{X}) \propto \tag{4.7}$$

$$L(\boldsymbol{X} \mid \boldsymbol{\pi}, \boldsymbol{\lambda}(\boldsymbol{.})_1, \ldots, \boldsymbol{\lambda}(\boldsymbol{.})_Q, \boldsymbol{\omega}) \times h(\boldsymbol{\pi}, \boldsymbol{\lambda}(\boldsymbol{.})_1, \ldots, \boldsymbol{\lambda}(\boldsymbol{.})_Q, \boldsymbol{\omega}).$$

The Bayesian way of parameter estimation is to obtain a sample from the posterior distribution (Gelman et al., 2000, p. 285-287; Schafer, 1997, p. 68-80; Zeger and Karim, 1991) and to calculate the EAP (Hoijtink, 2000) from this sample for each parameter. In this chapter we use a particular Markov Chain Monte Carlo (MCMC) method, that is Gibbs sampling, to obtain this sample. In Gibbs sampling the set of unknown parameters is split into a number of subsets. In each Gibbs iteration $z = 1, \ldots, Z$ each subset of parameters is sampled conditional on the most recently sampled values of all other subsets.

The structure of the resulting Gibbs sampling algorithm is the following (the reader is referred to Van Hattum and Hoijtink (2009b) for a detailed overview of the sampling algorithm):

1. To initialize the Gibbs sampler a reasonable allocation of the respondents to the latent clusters is needed. How this is done is explained in the next subsection. Furthermore, it is explained in this subsection how the Bayesian clustering approach handles large data sets. The missing values are initialized by setting them to 1.

2. For $q = 1, \ldots, Q$ and $j = 1, ..., J^0$ sample the cluster specific probabilities $\pi_{j|q}$ from $Post(\pi_{j|q}|\boldsymbol{x}_j^0, \boldsymbol{\tau})$. This is a $Dirichlet(\pi_{j|q} \mid N_{qj}^0 + 1, N_{qj}^1 + 1)$, where N_{qj}^0 denotes the number of respondents who did not pick item j and are currently allocated to cluster q and N_{qj}^1 denotes the number of respondents who did pick item j and are currently allocated to cluster q.

3. For $q = 1, \ldots, Q$ sample the cluster specific log-linear parameters $\boldsymbol{\lambda}(\boldsymbol{.})_q$ from $Post(\boldsymbol{\lambda}(\boldsymbol{.})_q \mid \boldsymbol{X}^*, \boldsymbol{\tau})$. This is achieved using Bayesian iterative proportional fitting (BIPF) (Gelman et al., 2000, p. 435-437; Schafer, 1997, p. 308-309). How BIPF works, is illustrated continuing the simple example from Section 4.2.2. The model in this example contains one intercept, three main effects and two two-way interactions.

   Let f_{abc}, for $a, b, c \in \{0, 1\}$ denotes the frequency with which each element of $\boldsymbol{Y}$ is observed in cluster q. Let θ_{abc} denotes the probability that a person in cluster q responds abc to the three items. To avoid heavy notation the subscript q is implicit for f_{abc} and θ_{abc}. Let $\boldsymbol{\theta} = (\theta_{000}, \ldots, \theta_{111})$.

   The first time the Gibbs sampler enters Step 3, $\theta_{abc} = \frac{1}{2^{J^*}}$, for $a, b, c \in \{0, 1\}$. In all other iterations Step 3 consists of two sub-steps:

   (a) Sample g_{ab+} from a standard Gamma distribution with shape parameters $f_{ab+} + 1$, for $a, b \in \{0, 1\}$, where $f_{ab+} = \sum_c f_{abc}$. Let $g_{+++} = \sum_{ab} g_{ab+}$, then,

   $$\theta_{abc}^{new} = \theta_{abc}^{current}(\frac{g_{ab+}/g_{+++}}{\theta_{ab+}^{current}}) \text{ , for } a, b, c \in \{0, 1\}. \tag{4.8}$$

   (b) Sample g_{a+c} from a standard Gamma distribution with shape parameters $f_{a+c} + 1$, for $a, c \in \{0, 1\}$, where $f_{a+c} = \sum_b f_{abc}$. Let $g_{+++} = \sum_{ac} g_{a+c}$, then

$$\theta^{new}_{abc} = \theta^{current}_{abc}(\frac{g_{a+c}/g_{+++}}{\theta^{current}_{a+c}}) \text{ , for } a, b, c \in \{0, 1\}. \tag{4.9}$$

After execution of these two sub-steps of Step 3 the parameters of the log-linear model are computed using

$$\boldsymbol{\lambda} = (\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T log\boldsymbol{\theta}. \tag{4.10}$$

(Schafer 1997, p. 299), where $\boldsymbol{\theta} = \{\theta_{000}, \ldots, \theta_{111}\}$.

4. Sample the cluster weights $\boldsymbol{\omega}$ from $Post(\boldsymbol{\omega} \mid \boldsymbol{\tau})$. This is a $Dirichlet(\omega_q \mid N_1 + 1, \ldots, N_q + 1, \ldots, N_Q + 1)$, where N_q denotes the number of respondents currently allocated to cluster q. See Narayanan (1990) for an overview of Dirichlet sampling methods.

5. For $i = 1, \ldots, N$ sample the respondents unobserved cluster memberships τ_i. This is a $Multinomial(\tau_i \mid \omega_{1|i}, \ldots, \omega_{q|i}, \ldots, \omega_{Q|i})$, where

$$\omega_{q|i} = \frac{\omega_q P(\boldsymbol{x}_i \mid \tau_i = q)}{\sum_{q'=1}^{Q} \omega_{q'} P(\boldsymbol{x}_i \mid \tau_i = q')}. \tag{4.11}$$

6. The missing values are handled by means of data augmentation (Hoijtink, 2000; Zeger and Karim, 1991). Sequentially, sample each element of the set $\{x_{ij} \mid m_{ij} = 0\}$. This is a *Bernouilli* distribution with a "success" probability that is calculated as follows:

   - if item j is a locally independent item, the "success" probability is

   $$P(x^0_{ij} = 1 \mid m_{ij} = 0, \boldsymbol{\pi}_q, \tau_i = q) = \pi_{j|q}, \tag{4.12}$$

   - if item j is a locally dependent item, the "success" probability is

   $$P(x^*_{ij} = 1 \mid m_{ij} = 0, \boldsymbol{\lambda}(.)_q, \tau_i = q) = \frac{exp^{\boldsymbol{R}_p\boldsymbol{\lambda}(.)_q}}{exp^{\boldsymbol{R}_p\boldsymbol{\lambda}(.)_q} + exp^{\boldsymbol{R}_s\boldsymbol{\lambda}(.)_q}}, \tag{4.13}$$

   where $\boldsymbol{R}_p$ is the row from $\boldsymbol{R}$ for which $y_{p1} = x_{i1}, \ldots, y_{pj} = 1, \ldots, y_{pJ^*} = x_{iJ^*}$ and $\boldsymbol{R}_s$ is the row from $\boldsymbol{R}$ for which $y_{s1} = x_{i1}, \ldots, y_{sj} = 0, \ldots, y_{sJ^*} = x_{iJ^*}$.

As is shown in Van Hattum and Hoijtink (2009b) the Gibbs sampling algorithm $Z = 1100$ with 100 burn-in iterations is sufficiently large to obtain convergence. In other words, the remaining sample of 1000 iterations constitutes a sample from the posterior (4.7).

Size of Data Sets

Using the hierarchical cluster algorithm described below and the Gibbs sampler from the previous subsection, Bayesian model based clustering can handle data sets, containing a large number of items and respondents.

Hoijtink and Notenboom (2004) present two conjectures with respect to the geometry of the posterior distribution of a standard cluster model, that are:

Conjecture 1. The mode of the posterior distribution for $Q - 1$ clusters equals the mode of the posterior distribution for Q clusters with two of the Q clusters combined into one. See Section 4.2.3 for an illustration of this conjecture.

Conjecture 2. Let Q_{max} denotes the maximum number of clusters for the data.

1. If $Q_{max} = 1$, sampling the posterior with $Q = 2$ renders the one cluster and an empty cluster.
2. If $Q_{max} = 2$, sampling the posterior with $Q = 2$ renders the two clusters.
3. If $Q_{max} > 2$, sampling the posterior with $Q = 2$ renders two clusters that are non-intersecting combinations of the Q_{max} clusters.

From these conjectures Hoijtink and Notenboom (2004) derive a hierarchical cluster algorithm that deals with large data sets and that ensures good initial allocations, as required in Step 1 of the Gibbs sampler.

Let Q_{max} denotes the maximum number of clusters for (a subset of) the data. In the very first iteration of the hierarchical algorithm the whole sample of respondents is randomly split into $Q = 2$ clusters. Subsequently the Gibbs sampler described in the previous subsection is applied to allocate each person to one of the two clusters.

In all subsequent iterations of the hierarchical algorithm, the algorithm:

1. determines which cluster is the largest of the Q clusters at hand,
2. randomly splits the respondents from the largest cluster into two clusters,
3. applies the Gibbs sampling algorithm to these two clusters only, to obtain a cluster membership for each person,
4. applies the Gibbs sampling algorithm to all $Q + 1$ current clusters to allow between cluster migration.

If the resulting number of respondents in each of the $Q + 1$ clusters is at least one, a new iteration is started. If at least one of the resulting $Q + 1$ clusters is empty, the current iteration is repeated by splitting the next largest cluster. If there is no next largest cluster left, the hierarchical algorithm stops. The result of the hierarchical algorithm is a sample from the global mode of the posterior distribution for Q_{max} clusters.

Number of Clusters

Note that according to the conjectures Q_{max} does not have to be pre-specified. It is an outcome of the hierarchical algorithm and an estimate of the number of clusters in the population from which the data are sampled.

Local Maxima

As mentioned in Section 4.2.3 clustering algorithms can converge on a local mode, or, in the Bayesian case, can converge on a local mode of the posterior distribution rather than converging on the globally optimal solution. However, Conjecture 1 defines local modes of the posterior distribution of the cluster model: a local mode consists of $Q < Q_{max}$ clusters, each corresponding to non-overlapping combinations of the Q_{max} clusters. This conjecture implies that there are no local modes for $Q = Q_{max}$ and that the global mode is reached when $Q = Q_{max}$.

That the hierarchical algorithm reaches the global mode is illustrated using the simulation study in Section 4.2.3. Analyzing the simulated data set with three different initializations render three different solutions with $Q = 2$ clusters. However, according to Conjecture 1 the hierarchical algorithm stops when the global maximum solution is reached. So with the $Q = 2$ cluster solutions the hierarchical algorithm does another iteration and finally reaches a $Q_{max} = 3$ cluster solution for each three initializations. As can be seen in Table 4.3, the three solutions, started from three different random initializations, are in close correspondence with the population from which the data were sampled (see Table 4.1).

Missing Values

As can been seen in Step 6, the sampling algorithm deals with missing values under the assumption that the missingness is missing at random (MAR) (Schafer and Graham, 2002). As a result all the respondents are assigned to one of the Q latent clusters.

Table 4.3: Parameter Estimates Obtained using the Hierarchical Algorithm from Three Different Random Initializations.

	Initialization 1			Initialization 2			Initialization 3		
Cluster q	1	2	3	1	2	3	1	2	3
ω_q	0.327	0.332	0.341	0.327	0.334	0.339	0.327	0.334	0.339
$\pi_{1\|q}$	0.83	0.10	0.08	0.83	0.10	0.08	0.83	0.10	0.08
$\pi_{2\|q}$	0.80	0.09	0.11	0.80	0.09	0.10	0.80	0.09	0.10
$\pi_{3\|q}$	0.82	0.08	0.09	0.83	0.09	0.09	0.82	0.08	0.09
$\pi_{4\|q}$	0.09	0.81	0.09	0.09	0.81	0.09	0.09	0.81	0.10
$\pi_{5\|q}$	0.11	0.80	0.14	0.11	0.80	0.14	0.11	0.80	0.14
$\pi_{6\|q}$	0.10	0.77	0.13	0.10	0.77	0.13	0.10	0.77	0.13
$\pi_{7\|q}$	0.10	0.11	0.80	0.10	0.11	0.80	0.10	0.11	0.80
$\pi_{8\|q}$	0.11	0.12	0.78	0.11	0.12	0.78	0.11	0.12	0.78
$\pi_{9\|q}$	0.10	0.10	0.81	0.10	0.10	0.82	0.10	0.10	0.81
$\pi_{10\|q}$	0.52	0.56	0.54	0.52	0.56	0.54	0.52	0.55	0.54

Local Item Dependence

The Bayesian model based approach has the possibility to incorporate locally dependent items. The probability $P(\boldsymbol{x}_i^* \mid \tau_i = q)$ is defined as in (4.5), where $\boldsymbol{\lambda}_q = \{\lambda_{0,q}, \lambda_{1,q}, \ldots, \lambda_{j,q}, \ldots, \lambda_{J^*,q}, \lambda_{\cdot,\cdot,q}, \ldots, \lambda_{\cdot,\cdot,q}\}$, for $q = 1, \ldots, Q$. Note that within each latent cluster the model contains an intercept, J^* main effects and (a subset of) the two-way interaction effects.

4.3 Evaluation

In order to evaluate the performance of the model based clustering approaches, we use simulated data sets. In the next subsections we use data sets that are simulated from a population in which the items are locally independent and locally dependent within the latent clusters. Furthermore, a large data set from a Dutch mail order company is described and it is discussed how the model based clustering approaches deal with this data set.

4.3.1 Performance Diagnostics

The performances of the model based clustering approaches is determined using the following performance diagnostics:

- Number of clusters: does the model based clustering approach render the same number of clusters as present in the population from which the data are sampled?

- Percentage of misclassifications: what is the fraction of estimated cluster memberships that does not match the simulated cluster memberships?

- Cluster specific parameters; do the estimated parameters, that are the cluster weights and the cluster specific probabilities, are in close correspondence with the parameters used to simulate the data?

- Runtime: how long does it take to render cluster specific parameters and cluster memberships?

To get the most reliable comparison, the model based clustering approaches are run on the same computer (Intel Core 2 CPU 2 GHZ.). Also the input parameters for LatentGold and Glimmix are kept equal, that is: the maximum number of EM-algorithms is set to 1000 and the EM-convergence criterion is the default setting, that is for LatentGold 0.01 and for Glimmix 0.00001. The convergence criterion in LatentGold is kept high, because the program switches to Newton Raphson after EM convergence. The convergence criterion in Glimmix is kept small because this program only uses EM. For both LatentGold and Glimmix the number of random initializations is set to 10 (this is the default setting in LatentGold, the default setting in Glimmix is one and therefore adjusted to 10).

In the case of the Bayesian approach the number of Gibbs iteration is set to 1100. Note that, as described in Section 4.2.6, these iterations apply to each split (Step 3) and each migration step (Step 4) in the hierarchical algorithm.

4.3.2 Local Item Independence within the Latent Clusters

The first simulated data set with items that are independent within each latent cluster, is described in Section 4.2.3. This data set contains 1,000 respondents with 10 locally independent items, split into $Q_{max} = 3$ clusters. The cluster specific parameters used to simulate the data set are shown in Table 4.1. Running the simu-

Table 4.4: Performance Diagnostics for Simulation 1

Performance diagnostics	Latent Gold	Glimmix	Bayesian approach
Number of clusters	3	3	3
Percentage misclassifications	4.1%	4.3%	4.4%
Cluster specific parameters	OK	OK	OK
Runtime	6 sec	3 hours	19 sec

lated data set with each model based clustering approach[‡] renders the performance diagnostics in Table 4.4.

From this table we can conclude that all model based clustering approaches are able to reconstruct the number of clusters used to simulate the data set. In terms of misclassifications Table 4.4 shows that the difference are rather small; LatentGold with 4.1% misclassifications, Glimmix with 4.3% misclassifications and the Bayesian approach with 4.4% misclassifications.

When looking to the cluster specific parameters in Table 4.5 it can be concluded that the estimated cluster weights and cluster specific probabilities in LatentGold, Glimmix and the Bayesian approach are in close correspondence with the parameters that are used to generate the data in Table 4.1. This explains the 'OK' in the third line of Table 4.4.

The biggest difference between the model based clustering approaches is the runtime. LatentGold is the fastest with 8 seconds, followed by the Bayesian approach with 19 seconds. The slowest is Glimmix with 3 hours. The reason is the small value of the EM convergence criterion (Section 4.3.1) in Glimmix.

The second simulated data set is a large data set containing 50,000 respondents with 156 locally independent items, split into $Q_{max} = 13$ clusters. The cluster specific parameters used to simulate the data set are shown in Table 4.6.

Running this large simulated data set is a problem in Glimmix. As described in Section 4.2.5, Glimmix is able to analyze data sets with up to 150 independent items and up to 50,000 respondents. Therefore we only analyze this simulated data set with LatentGold and the Bayesian approach.

Analysis of the simulated data set[§] renders the performance diagnostics, as displayed in Table 4.7. From this table we can conclude that both LatentGold and

[‡] for both LatentGold and Glimmix the range of cluster solutions is set to 1-5. After running these five cluster models, we used the BIC and CAIC to determine the number of clusters.

[§] for LatentGold the range of cluster solutions is set to 5-15. After running these 11 cluster models, we used the BIC and CAIC to determine the number of clusters.

Table 4.5: Cluster Specific Parameters for Simulation 1

	LatentGold			Glimmix			Bayesian approach		
Cluster q	1	2	3	1	2	3	1	2	3
ω_q	0.324	0.332	0.344	0.323	0.336	0.341	0.323	0.335	0.342
$\pi_{1\|q}$	0.85	0.10	0.06	0.85	0.10	0.07	0.85	0.10	0.06
$\pi_{2\|q}$	0.80	0.09	0.11	0.80	0.09	0.11	0.80	0.09	0.11
$\pi_{3\|q}$	0.84	0.08	0.08	0.84	0.08	0.08	0.84	0.08	0.08
$\pi_{4\|q}$	0.08	0.83	0.08	0.08	0.83	0.08	0.08	0.83	0.08
$\pi_{5\|q}$	0.11	0.80	0.15	0.11	0.79	0.14	0.11	0.79	0.15
$\pi_{6\|q}$	0.10	0.76	0.14	0.09	0.76	0.14	0.09	0.76	0.14
$\pi_{7\|q}$	0.09	0.10	0.80	0.09	0.11	0.80	0.09	0.10	0.80
$\pi_{8\|q}$	0.11	0.11	0.77	0.11	0.12	0.77	0.11	0.11	0.77
$\pi_{9\|q}$	0.09	0.09	0.82	0.09	0.09	0.83	0.09	0.09	0.83
$\pi_{10\|q}$	0.52	0.55	0.54	0.51	0.56	0.54	0.51	0.56	0.54

the Bayesian approach are not able to reconstruct the number of clusters used to simulate the data set. Both approaches render 14 clusters instead of 13 clusters. However, looking at the cluster weights in Table 4.8 for both approaches, it can be seen that the extra cluster is relatively small (cluster weight ω_{14} for LatentGold represents 66 respondents and cluster weight ω_{14} for the Bayesian approach represents 1 respondent). It can also be seen that the cluster weights for LatentGold and the Bayesian approach are both in close correspondence with the cluster weights used to simulate the data set at hand. The same holds for the cluster specific probabilities (not shown in a table), which explains the 'OK' in the third line of Table 4.7. In terms of misclassifications Table 4.7 shows that the Bayesian approach and LatentGold (0.7% vs. 0.6%) perform equally well.

The biggest difference between LatentGold and the Bayesian approach is the runtime. With LatentGold it takes about 140 hours to get cluster specific probabilities and cluster memberships. With the Bayesian approach it takes about 16 hours.

4.3.3 Local Item Dependence within the Latent Clusters

The third simulated data set is a data set with items that are locally dependent within the latent clusters. This data set contains 1000 respondents with 20 locally

Table 4.6: Cluster Specific Parameters for Simulation 2

	Population				
Cluster q	1	2	$\ldots$	12	13
ω_q	0.050	0.050	$\cdots$	0.050	0.050
$\pi_{1\|q}$	0.80	0.20	$\cdots$	0.20	0.20
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
$\pi_{12\|q}$	0.80	0.20	$\cdots$	0.20	0.20
$\pi_{13\|q}$	0.20	0.80	$\cdots$	0.20	0.20
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
$\pi_{24\|q}$	0.20	0.80	$\cdots$	0.20	0.20
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\pi_{133\|q}$	0.20	0.20	$\cdots$	0.80	0.20
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
$\pi_{144\|q}$	0.20	0.20	$\cdots$	0.80	0.20
$\pi_{145\|q}$	0.20	0.20	$\cdots$	0.20	0.80
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
$\pi_{156\|q}$	0.20	0.20	$\cdots$	0.20	0.80

Table 4.7: Performance Diagnostics for Simulation 2

Performance diagnostics	Latent Gold	Bayesian approach
Number of clusters	14	14
Percentage misclassifications	0.7%	0.6%
Cluster specific parameters	OK	OK
Runtime	140 hours	16 hours

Table 4.8: Cluster Weights ω_q for Simulation 2

	Population	Latent Gold	Bayesian approach
Cluster 1	0.051	0.051	0.051
Cluster 2	0.050	0.049	0.049
Cluster 3	0.101	0.101	0.101
Cluster 4	0.101	0.101	0.101
Cluster 5	0.103	0.103	0.103
Cluster 6	0.049	0.049	0.049
Cluster 7	0.047	0.047	0.047
Cluster 8	0.099	0.099	0.099
Cluster 9	0.099	0.099	0.099
Cluster 10	0.100	0.100	0.100
Cluster 11	0.101	0.101	0.101
Cluster 12	0.050	0.050	0.050
Cluster 13	0.050	0.048	0.049
Cluster 14	n.a.	0.001	0.000

dependent items and ten two-way interactions: $\lambda_{1,2}, \ldots, \lambda_{19,20}$. The data set is split into $Q_{max} = 5$ clusters. The cluster specific parameters used to simulate the data set are shown in Tables 4.9 and 4.10. Because it is not possible in Glimmix to model local item dependencies within the latent clusters, we only analyze the simulated data sets in this subsection with LatentGold and the Bayesian approach.

Analysis of the simulated data set[¶] renders the performance diagnostics in Table 4.11. From this table we can conclude that both LatentGold and the Bayesian approach are able to reconstruct the number of clusters used to simulate the data set. Furthermore, Table 4.11 shows that, according to the percentage of misclassifications, LatentGold (14.0%) and the Bayesian approach (14.7%) perform equally well.

Looking at the cluster specific parameters in Tables 4.9 and 4.10 it can be concluded that the estimated cluster weights and cluster specific probabilities in LatentGold and the Bayesian approach are in close correspondence with the parameters used to simulate the data set. This explains the 'OK' in the third line of Table 4.11.

[¶] for LatentGold the range of cluster solutions is set to 1-10. For each of the 10 cluster models the interaction effects are pre-specified. In other words, the local item dependencies within the clusters are pre-specified (which is also the case in the Bayesian approach). After running these 10 cluster models, we used the BIC and CAIC to determine the number of clusters.

Table 4.9: Cluster Specific Parameters for Simulation 3 (first part)

	Population						LatentGold						Bayesian approach				
Cluster q	1	2	3	4	5		1	2	3	4	5		1	2	3	4	5
ω_q	0.120	0.303	0.254	0.226	0.097		0.125	0.299	0.261	0.230	0.085		0.115	0.298	0.257	0.244	0.086
$\pi^{0,0}_{1,2\|q}$	0.22	0.44	0.07	0.06	0.29		0.29	0.47	0.06	0.03	0.33		0.27	0.47	0.06	0.04	0.34
$\pi^{0,1}_{1,2\|q}$	0.60	0.10	0.12	0.11	0.07		0.60	0.08	0.15	0.13	0.05		0.62	0.08	0.14	0.14	0.03
$\pi^{1,0}_{1,2\|q}$	0.07	0.41	0.43	0.42	0.57		0.06	0.42	0.44	0.38	0.60		0.06	0.42	0.44	0.37	0.60
$\pi^{1,1}_{1,2\|q}$	0.11	0.05	0.38	0.42	0.07		0.06	0.03	0.35	0.46	0.02		0.05	0.03	0.35	0.44	0.02
$\pi^{0,0}_{3,4\|q}$	0.30	0.31	0.40	0.11	0.61		0.34	0.31	0.37	0.16	0.65		0.37	0.31	0.35	0.17	0.64
$\pi^{0,1}_{3,4\|q}$	0.09	0.55	0.40	0.04	0.16		0.04	0.57	0.46	0.02	0.16		0.02	0.57	0.47	0.02	0.19
$\pi^{1,0}_{3,4\|q}$	0.43	0.04	0.08	0.57	0.17		0.43	0.04	0.08	0.59	0.15		0.38	0.04	0.08	0.59	0.14
$\pi^{1,1}_{3,4\|q}$	0.19	0.10	0.12	0.28	0.06		0.19	0.09	0.10	0.23	0.04		0.23	0.09	0.09	0.22	0.03
$\pi^{0,0}_{5,6\|q}$	0.58	0.32	0.16	0.08	0.32		0.74	0.29	0.13	0.07	0.35		0.72	0.32	0.12	0.11	0.30
$\pi^{0,1}_{5,6\|q}$	0.07	0.07	0.08	0.14	0.52		0.05	0.06	0.09	0.10	0.59		0.04	0.05	0.09	0.11	0.59
$\pi^{1,0}_{5,6\|q}$	0.30	0.48	0.45	0.23	0.05		0.18	0.51	0.51	0.18	0.02		0.20	0.50	0.52	0.17	0.07
$\pi^{1,1}_{5,6\|q}$	0.05	0.14	0.31	0.55	0.11		0.02	0.13	0.27	0.65	0.04		0.03	0.13	0.28	0.61	0.03
$\pi^{0,0}_{7,8\|q}$	0.42	0.52	0.29	0.78	0.27		0.49	0.50	0.32	0.78	0.19		0.49	0.49	0.32	0.79	0.19
$\pi^{0,1}_{7,8\|q}$	0.06	0.22	0.02	0.05	0.19		0.07	0.22	0.02	0.05	0.07		0.05	0.24	0.02	0.05	0.05
$\pi^{1,0}_{7,8\|q}$	0.31	0.09	0.52	0.12	0.14		0.26	0.12	0.48	0.11	0.21		0.28	0.12	0.48	0.10	0.22
$\pi^{1,1}_{7,8\|q}$	0.20	0.17	0.17	0.04	0.40		0.18	0.16	0.17	0.06	0.53		0.18	0.15	0.18	0.06	0.55
$\pi^{0,0}_{9,10\|q}$	0.08	0.02	0.04	0.13	0.18		0.06	0.01	0.03	0.14	0.15		0.06	0.01	0.03	0.14	0.16
$\pi^{0,1}_{9,10\|q}$	0.07	0.16	0.88	0.13	0.38		0.10	0.16	0.88	0.13	0.31		0.09	0.16	0.89	0.14	0.30
$\pi^{1,0}_{9,10\|q}$	0.82	0.60	0.04	0.70	0.41		0.82	0.59	0.06	0.71	0.54		0.85	0.59	0.05	0.69	0.53
$\pi^{1,1}_{9,10\|q}$	0.03	0.22	0.04	0.03	0.04		0.02	0.24	0.03	0.02	0.00		0.00	0.23	0.03	0.03	0.00

Table 4.10: Cluster Specific Parameters for Simulation 3 (second part)

	Population						LatentGold						Bayesian approach				
Cluster q	1	2	3	4	5		1	2	3	4	5		1	2	3	4	5
ω_q	0.120	0.303	0.254	0.226	0.097		0.125	0.299	0.261	0.230	0.085		0.115	0.298	0.257	0.244	0.086
$\pi^{0,0}_{11,12\|q}$	0.12	0.22	0.40	0.15	0.15		0.14	0.21	0.43	0.15	0.12		0.16	0.21	0.43	0.16	0.12
$\pi^{0,1}_{11,12\|q}$	0.53	0.16	0.33	0.12	0.04		0.53	0.13	0.31	0.08	0.00		0.57	0.13	0.31	0.08	0.00
$\pi^{1,0}_{11,12\|q}$	0.12	0.46	0.19	0.54	0.72		0.09	0.52	0.21	0.53	0.79		0.06	0.51	0.21	0.53	0.79
$\pi^{1,1}_{11,12\|q}$	0.23	0.15	0.07	0.20	0.09		0.24	0.14	0.05	0.23	0.09		0.21	0.16	0.05	0.23	0.09
$\pi^{0,0}_{13,14\|q}$	0.09	0.12	0.29	0.04	0.02		0.06	0.12	0.25	0.00	0.01		0.07	0.12	0.25	0.00	0.01
$\pi^{0,1}_{13,14\|q}$	0.22	0.12	0.30	0.26	0.71		0.19	0.14	0.28	0.31	0.75		0.15	0.14	0.27	0.34	0.76
$\pi^{1,0}_{13,14\|q}$	0.62	0.73	0.39	0.55	0.11		0.70	0.71	0.45	0.50	0.07		0.72	0.70	0.46	0.50	0.09
$\pi^{1,1}_{13,14\|q}$	0.07	0.03	0.02	0.16	0.17		0.05	0.02	0.02	0.18	0.16		0.06	0.03	0.02	0.16	0.14
$\pi^{0,0}_{15,16\|q}$	0.32	0.20	0.76	0.21	0.12		0.28	0.19	0.73	0.18	0.07		0.31	0.18	0.73	0.19	0.07
$\pi^{0,1}_{15,16\|q}$	0.09	0.10	0.10	0.34	0.43		0.09	0.11	0.10	0.36	0.41		0.07	0.11	0.11	0.35	0.43
$\pi^{1,0}_{15,16\|q}$	0.37	0.35	0.11	0.11	0.05		0.29	0.35	0.13	0.07	0.07		0.29	0.36	0.12	0.07	0.07
$\pi^{1,1}_{15,16\|q}$	0.22	0.35	0.03	0.34	0.40		0.34	0.35	0.04	0.39	0.45		0.33	0.35	0.04	0.39	0.43
$\pi^{0,0}_{17,18\|q}$	0.38	0.23	0.09	0.50	0.12		0.42	0.23	0.07	0.47	0.07		0.41	0.22	0.07	0.48	0.06
$\pi^{0,1}_{17,18\|q}$	0.18	0.08	0.49	0.12	0.03		0.17	0.05	0.57	0.12	0.06		0.17	0.04	0.57	0.12	0.08
$\pi^{1,0}_{17,18\|q}$	0.33	0.55	0.09	0.32	0.73		0.30	0.59	0.07	0.35	0.71		0.27	0.62	0.07	0.33	0.67
$\pi^{1,1}_{17,18\|q}$	0.11	0.13	0.33	0.05	0.12		0.12	0.13	0.30	0.07	0.16		0.15	0.12	0.29	0.07	0.19
$\pi^{0,0}_{19,20\|q}$	0.71	0.11	0.39	0.42	0.20		0.70	0.08	0.38	0.39	0.28		0.73	0.08	0.37	0.39	0.30
$\pi^{0,1}_{19,20\|q}$	0.05	0.09	0.28	0.13	0.07		0.06	0.07	0.30	0.15	0.04		0.07	0.07	0.30	0.14	0.03
$\pi^{1,0}_{19,20\|q}$	0.14	0.09	0.04	0.12	0.18		0.16	0.09	0.03	0.10	0.25		0.15	0.11	0.04	0.09	0.22
$\pi^{1,1}_{19,20\|q}$	0.09	0.70	0.28	0.33	0.55		0.07	0.76	0.29	0.37	0.44		0.05	0.74	0.29	0.39	0.44

Table 4.11: Performance Diagnostics for Simulation 3

Performance diagnostics	Latent Gold	Bayesian approach
Number of clusters	5	5
Percentage misclassifications	14.0%	14.7%
Cluster specific parameters	OK	OK
Runtime	1 min	6 min

Once again, the biggest difference between the model based clustering approaches is the runtime. LatentGold delivers the output in one minute, the Bayesian approach in six minutes.

4.3.4 Dutch Mail Order Company

The motivation for this chapter is that in day-to-day business often data sets are found that contain a large number of items and respondents. One such data set, from a Dutch mail order company, is used in this subsection. This mail order company, specialized in office, workplace and warehouse supplies, is active in the business-to-business market. In order to optimally use the available data, a segmentation study is performed using all the data from the years 2001 to 2005. The goal of this segmentation study is to find groups of businesses with (more or less) the same ordering behavior. Within these groups the mail order company wants to set-up (differentiated) marketing activities, like for example, differentiated promotions, cross or up-selling strategies, or loyalty programs.

The data, that are used in this application, come from the mail order company's ordering database. In this database the ordering behavior of 45,610 businesses is recorded for the years 2001 to 2005. In total the mail order company sells more than 50,000 products, classified into 399 product categories. Examples of these product categories are: dust-bins, cups, plates, protective clothing, tape, storage boxes, buckets, hat-racks, fire extinguishers, office desks/chairs/lamps, labellers, staples, tools, cupboards, bookcases, clocks, overhead projectors, cleaning equipments, archive accessories, white boards, black boards, safes, alarm systems, envelopes, telephone/fax machines, et cetera.

Given the large number of items only LatentGold and the Bayesian approach are able to analyze this data set (Glimmix is limited to analyze up to 150 items). For both approaches we specify a clustering model, assuming that the 399 items are locally independent within the latent clusters.

When running this large data set with LatentGold the first challenge is to specify the range of the number of clusters (Section 4.2.4). Because we do not know the actual number of clusters, we have to specify a broad range. For this data set the range chosen was 10 to 30. The program was stopped after 250 hours, when it was finished with the estimates of the parameters of a solution with 19 clusters.

Running the Bayesian algorithm renders 19 clusters in 28 hours. Or, in other words, 19 groups of businesses with more or less the same ordering behavior are found. Using the estimated cluster specific probabilities the cluster can be described. It goes beyond the scope of this chapter to fully describe the clusters.

Although LatentGold takes a lot of runtime to estimate the cluster specific probabilities for the 19 cluster solution, the probabilities are in close correspondence with the Bayesian approach. However, in order to do a good analysis, the information criterion for the 19 cluster solution needs to be compared with the criteria for the solutions with 20 or more clusters. The runtime for LatentGold would be longer than the above mentioned 250 hours.

4.4 Discussion

This chapter described three different model based clustering approaches: a Bayesian model based clustering approach and the two approaches implemented in LatentGold and Glimmix. Each clustering algorithm was described and its performance was evaluated using simulation studies. It was shown how each algorithm dealt with the main challenges in model based clustering, that are determining the number of clusters and protection against local maxima. Furthermore, it was described how the algorithms deal with data sets containing a large number of items and a large number of respondents, missing values and items that are locally dependent within the latent clusters.

For the three clustering approaches, the parameters are estimated in three different ways. Glimmix uses the EM algorithm, which renders a rather slow algorithm. In order to speed up parameter estimation LatentGold uses the EM algorithm followed by Newton Raphson. Finally, the Bayesian approach uses a Gibbs sampler embedded in a hierarchical algorithm.

Furthermore, Glimmix has a limitation on the number of items and number of respondents. The missing values handling is a sort of mean substitution for nominal items, which from past research (Schafer and Graham, 2002) is not found to be sufficient. And last, cluster models with items that are locally dependent within each latent cluster can not be analyzed.

Although the runtime when analyzing large data sets increased considerably, LatentGold has no limitations on the number of items and number of respondents in a data set. Furthermore, LatentGold offers the possibility to incorporate items that are locally dependent within the latent clusters.

The Bayesian approach, finally, has no limitations on the number of items and the number of respondents. From the simulation studies it became clear that, especially for large data sets, this approach had the fastest runtime.

In order to prevent against local maxima, in both Glimmix and LatentGold, the probability of finding the optimal solution is increased by initializing the EM algorithm with multiple random starting values. However, there is still a risk to converge on a local maximum solution. According to two conjectures (which are illustrated in a simulation study) the Bayesian approach can not converge on a local maximum solution.

With LatentGold and Glimmix it was a challenge to find the number op clusters. Both algorithms used the information criteria BIC and CAIC to select the number of clusters from a pre-specified range of numbers of clusters. The problem with pre-specifying this range is the possibility of finding the best solution within this range, whereas the global maximum (especially in large data sets) might be outside this range. In the Bayesian approach there is no need to pre-specify a range. An estimate of the number of clusters is an outcome of the algorithm.

Given the simulation studies and the accompanying performance diagnostics, it is clear that the Bayesian approach has advantages over LatentGold and Glimmix. The Bayesian approach is fast enough with small data sets and the fastest with large data sets. Furthermore, the problem of pre-specifying a range for the number of clusters is not an issue. An estimate of the number of clusters is an outcome of the hierarchical algorithm used. Finally, according to Conjecture 1 and Conjecture 2 the hierarchical algorithm stops when the global maximum solution is reached.

Chapter 5

The Proof of the Pudding is in the Eating *Data Fusion: An Application in Marketing*

Abstract[*]

Data fusion, or combining multiple data sets in one data set, is not a new concept. However, due to the increasing desire of differentiated direct marketing strategies, it is getting more popular in marketing. This chapter shows how marketing information can be fused to a company's customer database. Using a real marketing application, two traditional data fusion methods, that are polytomeous logistic regression and a nearest neighbor algorithm, are compared with two model based clustering approaches. Finally, the results are evaluated using internal and external criteria.

[*] This chapter is published as Van Hattum, P. and Hoijtink, H. (2008). The Proof of the Pudding is in the Eating. Data fusion: An Application in Marketing.*Journal of Database Marketing & Customer Strategy Management, 15(4)*, 267-284.

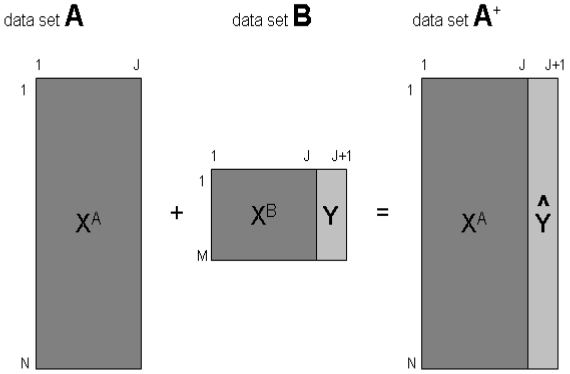


Figure 5.1: Schematic Representation of Data Fusion in Marketing (Derived and Adjusted from Van der Putten et al. (Van der Putten et al., 2002))

5.1 Introduction

In this chapter the following problem is addressed: a marketeer has knowledge and information about a small group of customers. Because the marketeer would like to have one-to-one communication with his customers, he would like to get the same knowledge and information for all the customers in his database. For reasons, like time, money, non-response, et cetera (Craig and McCann, 1978; D'Orazio et al., 2006; Kamakura and Wedel, 1997), obtaining the required knowledge and information using a single source questionnaire (Buck, 1989), is not an option. However, an attractive and practical solution is data fusion.

In this chapter data fusion is used in a marketing application. In the application a Dutch energy supplier wants to send differentiated questionnaires to all the customers in the database. However, for only a fraction of the customers it is known what kind of differentiated questionnaire is preferred. Using data fusion techniques, the information about the preferred differentiated questionnaire becomes known for all the customers in the database.

The general problem of data fusion can best be illustrated using the schematic representation in Figure 5.1. In this representation data set **A** is the customer database and contains knowledge and information (represented by J items) from all customers. Data set **B** contains knowledge and information (represented by $J + 1$ items) from a small group of customers. The first amount of knowledge and information (represented by the first J items) for a single customer is the same in each data set. However, from the small group of customers in data set **B** there is some additional knowledge and information, that is item $J + 1$. The goal of this chapter is to fuse the extra knowledge and information in data set **B**, that is item $J+1$, to data set **A**. As a result of this data fusion, the knowledge and information about item $J + 1$ becomes 'known' for all customers in the database, data set **A**.

Throughout the world different terminologies are used for above described fusion, or, integration of two (or more) data sets, for example: multi-source imputation, data attribution, data fusion, statistical record linkage, statistical matching, micro data set merging, et cetera (D'Orazio, 2006, p.2; Rässler, 2002, p. 2; Van der Putten et al., 2002). Since the 1980s a discussion has been going on about a clear and unambiguous terminology (Rässler, 2002, p. 2). As in European marketing literature and practice data fusion is the most commonly used term today (Gilula et al., 2006; Rässler, 2002, p. 2; Wedel and Kamakura, 2000, p. 256-257), the terminology of data fusion will be used in this chapter.

Not only is there some discussion about the terminology of integrating multiple data sets, there also is '*terminological confusion*' (D'Orazio, 2006, p.2) about the different (statistical) procedures of data set integration. The focus of this chapter is on integrating (or fusing) one single categorical item to another data set, whereas other papers (D'Orazio et al., 2006, p.3-5; Kamakura and Wedel, 1997; Rässler, 2002, p. 2-3; Van der Putten, 2002) focus on integrating (or fusing) multiple (categorical) items.

The structure of this chapter is as follows. Section 5.2 describes the concept of data fusion and how data fusion can be used in the context of marketing. This section also describes two more traditional algorithms, that are nearest neighbor methods (Dillon et al., 1978; Gilula et al., 2006; Rodgers, 1984) and polytomeous logistic regression methods (Hosmer and Lemeshow, 2000, Chapter 8.1), versus two newly made data fusion algorithms, that are methods based on model based clustering (Hoijtink and Notenboom, 2004), which are used in this chapter. Finally, this section shows how the four data fusion methods can be evaluated using internal and external validation criteria. This section also explains why we choose to work with real data sets, rather than simulated data sets, in order to validate the four data fusion algorithms. Section 5.2 illustrates data fusion with the use of the marketing application about the Dutch energy supplier. For the application the marketing goals are described, how data fusion is used to obtain the marketing goals and what the results are of applying the proposed data fusion algorithms. This chapter concludes with a discussion in Section 5.3.

5.2 Data Fusion

5.2.1 Introduction

The concept of data fusion is not new. Although there has always been resistance to do data fusion (Baker et al., 1989; Bronner, 1988; Rässler, 2002, p.1), there has been a great diversity in data fusion applications since the 1960s (see D'Orazio (2006, Chapter 3) and Rässler (2002, Chapter 7) for an overview of applications in Europe and the United States). It is since the 1980s that data fusion is also used for marketing purposes (D'Orazio, 2006, p.174; Gilula et al., 2006; Kamakura and Wedel, 1997). The most commonly used data fusion method in these applications, is based on nearest neighbor methods.

Section 5.2.2 shows how data fusion can be used for differentiated marketing purposes. Section 5.2.3 describes the two existing and the two new data fusion methods, that will be evaluated in this chapter. Section 5.2.4 explains why we prefer to use real data sets and cross-validation over simulated data sets, in order to evaluate the performance of the fusion methods under consideration in this chapter. Finally, Section 5.2.5 shows how data fusion procedures can be evaluated using two validation criteria.

5.2.2 Data Fusion in Marketing

'*Differentiated marketing builds greater loyalty and repeat purchasing by considering customer needs and wants. Differentiated marketing creates more total sales with a concentrated marketing effort in selected areas. Concentrated or target marketing gains market position with specialized market segments. Target marketing of products or services reduces the cost of production, distribution and promotion.*' (Bull and Passewitz, 1994). It is because of these benefits that differentiated marketing is getting more popular (Cui and Choudhury, 2002, 2003; Buckinx, 2005; Van der Putten et al., 2002). Instead of targeting customers with the same marketing strategy, companies want to target customers as individually as possible. Or, in other words, the company may be trying to sell exactly the same product or service, but it will change, for example, its promotional methods for (a group of) individuals.

In order to target (groups of) customers individually, it is important to know how they react on different marketing mix strategies. How do customers want their products or services to be packed? Where do they shop? Do they read advertisements? How do they react on discounts? All interesting facts companies need to know about their customers in order to set up good direct marketing strategies.

Information about customers can be found everywhere. An example is a company's customer database. Also market research is a powerful tool to get information about customers. In the past years these sources of customer data have grown exponentially (Van der Putten et al., 2002). It seems that collecting all the desired customer information in one single source market research questionnaire (Buck, 1989) is the best solution. But as time and money (D'Orazio et al., 2006; Kamakura and Wedel, 1997) is limited in most marketing companies, this is often not realized. An attractive and practical solution is data fusion, or, in other words, integrating different data sets.

Data fusion is used in the following marketing application. A Dutch energy supplier wants to know their customer's interests in energy products and services; what is their interest in: information about energy savings, solar panels, custom-made advice, government grants for energy, et cetera. The energy supplier wants to send a questionnaire to all their 1,133,405 customers in their customer database (data set **A** in Figure 5.1) in order to obtain the desired knowledge about their customers.

To get the highest response, the energy supplier decides to send differentiated written questionnaires. From past experiences the supplier knows what the responses are with regular (or undifferentiated) written questionnaires. Using the differentiated questionnaires the supplier hopes to trigger the interest of the customers and, consequently, to improve the response. Furthermore, from past experiences, the energy supplier knows, on average, in how many energy products and services, customers are interested. Using the differentiated approach the supplier hopes that the interests in energy products and services will increase (Feinberg et al., 2000; Lattin and Bucklin, 1989).

The energy supplier knows that each individual has a different attitude towards energy and issues related to energy. Because of this, a motivational research study, called Brand Strategy Research (BSR) [†] (Brethouwer et al., 1995, p. 8; Oppenhuisen, 2000, p. 79-81; Van Hattum and Hoijtink, 2009b), is conducted among 1,751 customers (data set **B** in Figure 5.1). The 1,751 customers are a fraction (simply using the whole customer database is too expensive) of the supplier's customer database. From the motivational study it is known that there are actually five groups (or clusters) of customers who have more of less the same attitude towards energy and issues related to energy. Short descriptions (see `www.smartagent.nl`

[†] BSR is based on Adler's social-psychology theory (Callebaut et al., 1999, p. 55-60) and provides a framework for understanding customers at the 'deepest' level. This motivational level gives knowledge of consumer's fears, beliefs and values, thus providing a understanding of the fundamental motivations that drive (future) purchase decisions of customers. The interested reader is referred to `www.smartagent.nl` for more information about BSR.

for detailed descriptions) of these five motivational clusters are:

- cluster 1: energy stands for creating a cozy and warm atmosphere. Customers in this cluster try to find a balance between their own comfort and the comfort for persons in their neighborhood. The usage of energy is a well-considered choice;
- cluster 2: for customers in this cluster energy is self-evident; the goal of the energy supplier must be to deliver as much energy as needed. Customers in this cluster are followers; the usage of customers from this cluster is mainly oriented on their peer group and the rules and values of this group;
- cluster 3: customers in this cluster use as much energy as needed for their own well-being and their own comfort; they do not conform to rules and values in society. Energy is a uncomplicated and single product. As such the energy supplier must deliver energy with a price as low as possible and with as least contact moments as possible;
- cluster 4: customers in this cluster feel guilty towards nature when using energy. The usage of energy is a well-considered choice. Customers from this cluster are looking for an energy supplier that is active in the field of energy saving technique;
- cluster 5: customers in this cluster think they are smarter than their energy supplier. They live according to their own (superior) rules and values. Customers in this cluster are looking for an energy supplier that acknowledge the customer's expertise in the field of energy. The usage of energy is a smart and well-considered choice. This results in all kind of energy saving products and services.

These five motivational clusters provide a basis for developing a company's vision and/or a company's marketing directions on the strategic, tactical and operational levels, aligning the total marketing mix around the consumers needs in the domain energy. Table 5.1 displays the frequencies of the resulting motivational clustering.

Using the descriptions of the five motivational clusters, for each cluster a separate questionnaire is made by a specialized communication agency. The content of the questionnaires, that are the questions about the customer's interests in energy products and services, is the same for each questionnaire. Only the lay out (colors and pictures used in the questionnaire) and the tone-of-voice of the invitation letters are different for the cluster specific questionnaires.

Because the energy supplier wants to send a differentiated written questionnaire to all their customers, data fusion is used. Using the fraction of the supplier's

Table 5.1: Frequency Respondents in Motivational Research Study Energy (Between Brackets are the Percentages Based on the Total Number of Respondents Classified to One of Five Motivational Clusters)

Cluster	Frequency respondents	Percentage respondents
Cluster 1	537	30.7% (31.7%)
Cluster 2	337	19.2% (19.9%)
Cluster 3	265	15.1% (15.6%)
Cluster 4	305	17.4% (18.0%)
Cluster 5	251	14.3% (14.8%)
No cluster	56	3.2%
Total	1,751	100.0%

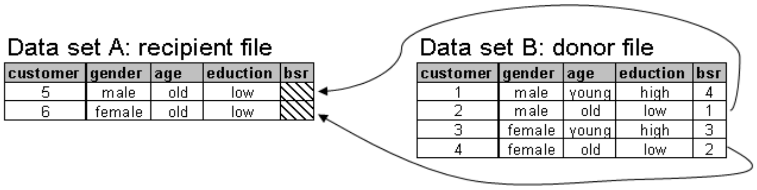
customer database (data set **B** in Figure 5.1) for which the motivational clusters are known, data fusion methods are used to fuse the motivational clusters to the rest of the supplier's customer database. In order to do this, ten items (data set **A** in Figure 5.1), that are gender, age, education, position in household, type of work, occupation, number of persons in household, household stage, social economic status and income, which are known for all the 1,133,405 customers, are used.

5.2.3 Methods and Algorithms Used

In literature data fusion problems have been solved by several, more traditional methods, like for example: regression techniques, discriminant analysis, nearest neighbor algorithms, network approaches, et cetera. More recently Kamakura and Wedel (1997) proposed a mixture model based methodology for data fusion. Each method with its own advantages and disadvantages.

But despite of all these advantages and disadvantages not all the above mentioned methods are appropriate for data fusion, as described in Section 5.2.2. In the following subsections two of above mentioned traditional methods for data fusion, that are nearest neighbor algorithms and regression techniques, are adjusted and described for the purpose of this chapter. Furthermore, two new methods based on model based clustering, are introduced and described.

In order to describe the data fusion methods, the schematic representation of data fusion in Figure 5.1 is used with the following notation: data set **A** is a customer database and contains information from $i = 1, \ldots, N$ customers about $j = 1, \ldots, J$ items ($\mathbf{X}$), where $\mathbf{X} = (x_{i1}, \ldots, x_{iJ})$, for all customer i. Data set **B**



Data set A: recipient file

customer	gender	age	eduction	bsr
5	male	old	low	▨
6	female	old	low	▨

Data set B: donor file

customer	gender	age	eduction	bsr
1	male	young	high	4
2	male	old	low	1
3	female	young	high	3
4	female	old	low	2

Figure 5.2: Nearest neighbor method

comes from a market research study and contains information from M customers about $J + 1$ items. The description of the first J items is equal for both data sets. The goal of this chapter is to fuse the extra information in data set $\mathbf{B}$, that is item $J+1$ ($\mathbf{Y}$), to data set $\mathbf{A}$. Where $\mathbf{Y} = y_i$, for all customer i. As a result of this data fusion the information about item $J + 1$ becomes 'known' for the N customers in data set $\mathbf{A}$, that is $\widehat{\mathbf{Y}}$ in data set $\mathbf{A}^+$.

Nearest Neighbor Method

In practice the most commonly used algorithms for data fusion are based on nearest neighbor methods (Dillon et al., 1978; Gilula et al., 2006; Rodgers, 1984; Wind and Mahajan, 1997). In other words, values that are missing in one data set, usually called the recipient data set, are duplicated from another data set, usually called donor data set. The choice of the duplication record from the donor data set is based on a certain (distance) measure, calculated on the common items in both data sets.

Translated to the application about the Dutch energy supplier, data set $\mathbf{A}$, the total customer database, is the recipient data set and data set $\mathbf{B}$, the fraction of the customer database containing data set $\mathbf{Y}$, is the donor data set. As is illustrated (for simplification only the first three items are used) in Figure 5.2, the motivational clusters in the recipient data set (column 'bsr'; bsr represents the resulting motivational clusters as described in Section 5.2.2.) are missing and needs to be fused using the records from the donor data set. As can be seen from this figure, customer record 5 in the recipient data set is exactly the same as customer record 2 in the donor data set. Consequently, the value of the motivational cluster in customer record 2 (bsr=1) in the donor data set is duplicated (or fused) to customer record 5 in the recipient data set. Likewise, the value of the motivational cluster in customer record 4 (bsr=2) in the donor data set is duplicated (or fused) to customer record 6 in the recipient data set.

An important aspect in nearest neighbor algorithms is the choice of the (distance) measure, calculated on the common items in both data sets. Different measures have been used for data fusion, for example: Euclidean distance, City-block distance, Mahalanobis distance, et cetera. See D'Orazio et al. (2006) for calculation of several distances. Besides selecting the appropriate distance measure, the duplication of records can also be restricted by all kind of constraints. For example, girls less than 12 years old can't be pregnant, et cetera. See D'Orazio et al. (2006) and Rässler (2002) for an overview of different (un)constrained measures that can be used in nearest neighbor algorithms.

Despite the popularity of nearest neighbor methods in data fusion practice, the major disadvantage of these algorithms is the heuristic rule from which the duplication of data from the donor data is based. Kamakura and Wedel (1997) state that the choice of the type of distance measure is subjective and can critically affect the quality of the data fusion. Also D'Orazio et al. (2006) warn for these disadvantages when using nearest neighbor methods.

However, in practice, nearest neighbor methods are still the most commonly used in data fusion problems (Van der Putten et al., 2002; Wedel and Kamakura, 2000, p.256-257). In Germany, it is common practise to use Euclidean or City-Block distances (Rässler, 2002, p.18), where D'Orazio et al. (2006) state, that the Mahalanobis distance is the most popular distance in data fusion practise. This chapter uses a nearest neighbor method with an Euclidean distance measure.

Polytomeous Logistic Regression

Regression methods have become an important aspect of any data analysis. These methods are used to describe the relationship between an outcome item and some explanatory items. When the outcome item is categorical, logistic regression has become the standard method of analysis in this situation (Hosmer and Lemeshow, 2000, p.1). The best known usage of logistic regression, is the case in which the categorical outcome item has only two categories. In literature this is often called binary logistic regression. For an overview of binary logistic regression, see Hosmer and Lemeshow (2000, Chapter 3). According to Ratner (2003, Chapter 9.2), when the categorical outcome item has three or more levels, this is called polytomeous logistic regression. For an overview of polytomeous logistic regression, see Hosmer and Lemeshow (2000, Chapter 8.1) and Ratner (2003, Chapter 9.2).

Above description of logistic regression can also be used in data fusion (Bucklin and Gupta, 1992; Jones and Zufryden, 1980; Mela et al., 1997). This application of logistic regression techniques in data fusion problems is in fact a single imputation in a missing value problem (Little and Rubin, 2002, p.59; Schafer, 1997, p.197), that is using the estimates of the logistic regression model, fitted on the complete data set, the missing value, in this case the fusion value, is imputed for the incomplete data set.

Translated to the application about the Dutch energy supplier, a polytomeous logistic regression model is fitted, in order to describe the relationship between the motivational clusters in data set $\mathbf{Y}$ and the ten socio-economical items in data set $\mathbf{X}$ (in data set $\mathbf{B}$). Using the fitted regression coefficients, for each customer $i = 1, \ldots, N$ in data set $\mathbf{A}$, it is now possible to calculated the probabilities $P(y_i = 1 \mid x_{i1}, \ldots, x_{i10}), \ldots, P(y_i = 5 \mid x_{i1}, \ldots, x_{i10})$. Where customer i is classified to the motivational cluster value with the highest probability.

For all the technical details of fitting a logistic model, the reader is referred to Hosmer and Lemeshow (2000, p.262-264). This chapter uses the statistical software program SPSS 15.0 to fit the polytomeous logistic model.

Fusion Value Specific Probabilities Model

This new method is based on latent class analysis, where the role of the latent classes is taken by the fusion values and the explanatory variables are the items. This is illustrated in two steps, using a simple example (for simplification only the first three items, are used) in the context of the application about the Dutch energy supplier. As can be seen from Table 5.2 the items x_1 (gender), x_2 (age) and x_3 (education) in data set $\mathbf{B}$ are used to fit a fusion value specific probabilities model in order to predict (or fuse) item y (motivational cluster) to data set $\mathbf{A}$.

Step 1: Using data set $\mathbf{B}$ the fusion value sizes (displayed in the row 'ω_y' in Table 5.2) and the fusion value specific probabilities (displayed in column '1', '2', '3', '4' and '5' in Table 5.2) are estimated. For example, there are 263 customers classified to motivational cluster 1 (=fusion value 1), that is 0.30 of the total number of customers in the data set. From these 263 customers classified to motivational cluster 1, there are 105 customers for which $x_1 = 1$ and 158 customers for $x_1 = 2$. This is 0.40 and 0.60, respectively. Likewise, the other model parameters are calculated and displayed in Table 5.2).

Step 2: Using the classification rule of latent class analysis (Vermunt and Magidson, 2000, p. 148), the model parameters in Table 5.2 are used for fusing the motivational clusters to the customers in data set A. This is done in the following way: suppose a customer in data set A has the following answer pattern: $x_1 = 1$ (gender=male), $x_2 = 2$ (age=old) and $x_3 = 2$ (education=high). Using the estimated fusion value specific probabilities (column '1', '2', '3', '4' and '5') and the estimated fusion value sizes (row 'ω_y') the probabilities of fusing the motivational clusters to this customer, with answer pattern $x_1 = 1$ (gen-

Table 5.2: Model Parameters for the Fusion Value Specific Probabilities Approach. Note: these Counts and Probabilities are Fictive Figures.

y		1	2	3	4	5	total
ω_y		0.30	0.20	0.16	0.18	0.17	1.00
		(263)	(173)	(139)	(155)	(145)	(875)
x_1	1	0.40	0.20	0.44	0.52	0.52	0.42
		(105)	(34)	(61)	(80)	(75)	(365)
	2	0.60	0.80	0.56	0.48	0.48	0.58
		(158)	(139)	(78)	(75)	(70)	(510)
x_2	1	0.43	0.80	0.46	0.74	0.32	0.46
		(114)	(139)	(64)	(115)	(46)	(399)
	2	0.57	0.20	0.54	0.26	0.68	0.54
		(149)	(34)	(75)	(40)	(99)	(476)
x_3	1	0.73	0.28	0.56	0.45	0.13	0.42
		(193)	(48)	(78)	(69)	(19)	(365)
	2	0.27	0.72	0.44	0.55	0.87	0.58
		(70)	(125)	(61)	(86)	(126)	(510)

der=male), $x_2 = 2$ (age=old) and $x_3 = 2$ (education=high), are calculated, that are $P(y = 1 \mid x_1 = 1, x_2 = 2, x_3 = 2) = 0.17$, $P(y = 2 \mid x_1 = 1, x_2 = 2, x_3 = 2) = 0.05$, $P(y = 3 \mid x_1 = 1, x_2 = 2, x_3 = 2) = 0.16$, $P(y = 4 \mid x_1 = 1, x_2 = 2, x_3 = 2) = 0.13$ and $P(y = 5 \mid x_1 = 1, x_2 = 2, x_3 = 2) = 0.49$. Because the probability of fusing motivational cluster 5 is the highest of the five probabilities, motivational cluster 5 is fused to this customer in data set **A**, with this particular answer pattern.

Model Based Clustering Approach

In recent years model based clustering has become a popular technique. Also in marketing, model based clustering has become an established tool (Kamakura and Wedel, 1997; Kamakura et al., 2003; Varki and Chintagunta, 2004). An important difference between traditional clustering (Hair et al., 1984, p. 469-518) and model based clustering (Banfield and Raftery, 1993; Bensmail et al., 1997; Fraley and Raftery, 1998; Kamakura and Wedel, 1997; Newcomb, 1886; Pearson, 1894; Varki and Chintagunta, 2004; Vermunt and Magidson, 2000, p. 1-2, 152) is that in the latter it is assumed that the data are generated by a certain mixture of underlying probability distributions. Kamakura and Wedel (1997), Moustaki and Papageorgiou (2005) and Vermunt and Magidson (2000, p. 2) describe some advantages of a probabilistic clustering approach.

A model based clustering approach has been developed, that can be used for data fusion in the context of this chapter. The goal of this model based clustering approach is to 'unmix' the mixture of underlying probability distributions. Translated to this chapter, the goal of the proposed model based clustering approach is to 'unmix' the fusion value specific probabilities from the previous subsection. As a result of the model based clustering approach, there will be a fusion value specific probabilities model for each latent cluster found. Translated to the application about the energy supplier, the number of latent clusters found is 16; for each of the 16 latent clusters a fusion value specific probabilities model is estimated.

This chapter does not go into detail about the model based clustering approach. The interested reader is referred to Hoijtink and Notenboom (2004) for all the technical details about the proposed model based clustering approach.

5.2.4 Data Sets: Real or Simulated?

The purpose of this chapter is to evaluate the data fusion methods introduced in the previous section. One of the most important evaluation criteria in comparing the four methods is the quality, or, reconstruction, of the individual (missing) values.

In order to show the number of mismatches for each fusion method, we need two thing. First of all, we need a training data set to which each of the four fusion models can be fitted. Secondly, we need a test data set with the true individual fusion values known, for which the predicted values are obtained using the fitted models. Comparing the true fusion values with the predicted values for each of the four fusion methods, gives us insight in the performances of the fusion methods. However, the problem in reality is the lack of test data sets with known fusion values.

A solution to above problem is simulating the training and the test data set. A major disadvantage of simulating data sets is that you can choose the simulation model (e.g. nearest neighbor, regression, model based clustering, et cetera) and the simulation parameters (e.g. regression parameters, number of clusters, within cluster parameters, et cetera), such that it favors one of the four data fusion methods. Another disadvantage of simulating data sets is, that it is almost impossible to choose the simulation model and the simulation parameters, such that the simulated data set is a good representation of reality. And more important, with simulated data sets it is impossible to validate the results of the data fusion externally (this is further described in the next subsection).

A good alternative for simulating data sets without the disadvantages of simulation is cross-validation (Kamakura et al., 2003; Verstraeten, 2005). In cross-validation the data set that is used for fitting the data fusion models and for determining where the true individual values are known, is randomly split into a training data set and test data set. The training data set is used for training (or calibrating) the data fusion models and, since the true individual values are known, the test data set is used for evaluating the fusion models. The use of cross-validation in the validation of the fusion models is further described in the next subsection.

In this chapter cross validation on a real data set is used in both marketing applications. Not favoring simulation models or simulation parameters in simulated data sets; the experiments described in this chapter are performed in their most realistic context.

5.2.5 Validation

After fusing two data sets, the big question is how good (or bad) is the data fusion. In her book Rässler (2002, p.29-30) describes four levels of data fusion validation. Rässler (2002, p.30) states that the first level of validation, that is the preservation of individual values, or, the reconstruction of the individual values, is the most challenging level of the data fusion validation. Furthermore, Rässler (2002, p.32) states that this first level is very difficult to achieve and in many case not practical. However, this is not the case in the context of this chapter.

In this chapter the goal is to fuse a data set, containing an item with information about a customer's reaction on a certain marketing mix strategy, to another data set. Taking this goal into consideration, it is undesirable that the reconstruction of customer's individual values are ill-performed. Or, translated to the application about the Dutch energy supplier, it is undesirable that a customer, that belongs to motivational cluster 1, is fused to motivational cluster 2. In order to show the realistic number of such mismatches, this chapter concentrates on Rässler's first level of data fusion validation. More specific, this chapter concentrates on both a validation step within the data set and on a validation step in the actual market, after a real marketing strategy has taken place. In this chapter the first validation step is called the internal validation. As described in the previous subsection, the internal validation step uses cross-validation for validation of the results. The second validation is called the external validation.

Internal Validation

One of the most important goals of this chapter is to minimize the number of mismatches, or, to maximize the number of correct matches, in the reconstruction of customer's individual values. Since the customer's true fusion values are not known, Rässler (2002, p.30) only validates by means of simulation studies. However, this chapter makes use of real data sets. In order to get an idea of the number of correct matches, data set $\mathbf{B}$ is randomly split according to a $2:1:1$ proportion. Which means that roughly $\frac{2}{4}$th of data set $\mathbf{B}$, or data set $\mathbf{B}_{train}$, is used for training (or calibrating) the data fusion model, roughly $\frac{1}{4}$th of data set $\mathbf{B}$, or, data set $\mathbf{B}^1_{test}$, is used for the first validation of the data fusion model and roughly $\frac{1}{4}$th of data set $\mathbf{B}$, or, data set $\mathbf{B}^2_{test}$, is used for the second validation of the data fusion model. In the case of the application about the Dutch energy supplier the number of customers in $\mathbf{B}_{train} = 875$, in $\mathbf{B}^1_{test} = 411$ and in $\mathbf{B}^2_{test} = 409$. Because the customer's true fusion values are known in the test data sets, the number of correct matches can easily be determined (Kamakura et al., 2003).

The advantage of splitting the data set $\mathbf{B}$ into a train data set and test data sets, is the prevention against model overfitting. Overfitting refers to the phenomenon in which a data fusion model may well describe the relationship between explanatory items and an outcome item in the data set used to develop the model, but may subsequently fail to provide valid predictions when cross validating a new data set. The model shows an adequate fit in the data set under study, but does not cross validate, that is, does not provide accurate predictions for observations from a new data set. In the remainder of this subsection some examples of model overfitting are shown. However, this chapter does not go into detail about this topic. The interested reader is referred to Verstraeten (2005) for more details of model overfitting. Two test data sets are used because of the dependency of the validation results of one particular split of the data set used (Verstraeten, 2005).

In order to draw conclusions about the quality of the data fusion, Ratner (2003, p.181-183) introduces the statistics model lift and total correct classification rate (TCCR). These statistics are explained and described using the application about the Dutch energy supplier. Table 5.3 displays the classification table after the data fusion method 'logistic regression' is applied on $\mathbf{B}_{train}$. As is described in Section 5.2.2, the fusing item is the motivational cluster about the domain energy. Customers are classified to either motivational cluster 1, 2, 3, 4 or 5. The row totals of Table 5.3 show the actual counts in data set $\mathbf{B}_{train}$. The column totals show how the predicted classification counts are after applying the data fusion method 'logistic regression' on data set $\mathbf{B}_{train}$. The percentages under the total counts (between brackets) are with respect to the total number of customers in data set $\mathbf{B}_{train}$. For example, in data set $\mathbf{B}_{train}$ the actual percentage of customers classified to motivational cluster 2 is 19.8%. However, the predicted percentage is 21.5%.

The diagonal in Table 5.3 displays the numbers of correct matches for each motivational cluster. For example, 90 customers, which is 47.9% $(=\frac{90}{188})$, are correct classified to motivational cluster 2. In this chapter this is called the total correct classification rate for motivational cluster 2 ($TCCR(2)$), which is derived from Rat-

Table 5.3: Classification Table

		PREDICTED					
		cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	total
ACTUAL	cluster 1	167 (50.9%)	43	18	21	14	263 (30.1%)
	cluster 2	61	90 (47.9%)	8	12	19	173 (19.8%)
	cluster 3	27	20	61 (48.4%)	12	19	139 (15.9%)
	cluster 4	35	25	14	59 (50.4%)	22	155 (17.7%)
	cluster 5	38	10	25	13	59 (50.9%)	145 (16.6%)
	total	328 (37.5%)	188 (21.5%)	126 (14.4%)	117 (13.4%)	116 (13.3%)	875 (100.0%)

ner's total correct classification rate for the overall model ($TCCR(model)$). However, using the actual percentages one would expect to find 19.8% of the customers to be classified to motivational cluster 2, or, in other words, based on a random chance model one would expect to find 19.8% of the customers to be classified to motivational cluster 2. In this chapter this is called the total correct classification rate for motivational cluster 2 that can be obtained by a random chance model ($TCCR_{chance}(2)$). Using these TCCR's the model lift for motivational cluster 2 is $\frac{TCCR(2)}{TCCR_{chance}(2)}$=242, which means that the data fusion method 'logistic regression' provides 142% more correct matches for motivational cluster 2 than obtained by chance.

These statistics can also be calculated in order to draw conclusions about the overall quality of the data fusion. From Table 5.3 it is clear that 436 (=167+90+61+59+59) customers are correctly classified to one of the five motivational clusters. This results in a total correct classification rate for the overall model ($TCCR(total)$) of 49.8% (=$\frac{436}{875}$). To calculate the model lift for the overall model, the $TCCR(total)$ is compared with the $TCCR_{chance}(total)$, that is the total correct classification rate for the overall model that can be obtained by a random chance model. The $TCCR_{chance}(total)$ is defined as the sum of the square actual value percentages. For Table 5.3 the $TCCR_{chance}(total)$ is 21.4% (=30.1%^{2}+19.8%^{2}+15.9%^{2}+17.7%^{2}+

16.6%^{2}). Using these TCCR's the overall model lift is $\frac{TCCR(total)}{TCCR_{chance}(total)}$=233, which means that the data fusion model provides 133% more correct matches for all the motivational clusters than obtained by chance.

Above described classification table are made for each data fusion method, applied to one of the tree data sets. However, the figures that are the most important from these tables, are the actual and predicted frequencies and the information necessary to calculate the statistics TCCRs and model lifts. Table 5.4, 5.5 and 5.6 summarize this important information. Table 5.4 displays the actual and predicted frequencies after applying the data fusion methods to the train and two data sets. This table also shows how many customers are in each data set. Table 5.5 displays the total correct classification rates (TCCRs) for each motivational cluster and for the total model. This table also shows what the percentage of correct matches would be in each data set when obtained by chance. Table 5.6 displays the model lifts for each motivational cluster and for the total model. Note that in all three tables the figures in the nearest neighbor method for the train data set are missing. This because the train data set is defined as the donor data set (see Section 5.2.3) and the motivational clusters are duplicated from this data set for the two test data sets (the recipients files).

In order to determine which data fusion method performs the best, several considerations need to be made. First of all, for each data fusion method, the predicted frequencies of the motivational clusters are compared with the actual frequencies. Table 5.4 shows that, for each data fusion method, the predicted frequencies for motivational cluster 2, 3 and 5 are closer to the actual frequencies. The predicted frequencies for motivational cluster 1 and 4 are more different.

Secondly, the TCCRs and the model lifts are examined for each data fusion method. Corresponding with this second consideration, a third consideration, the degree of model overfitting plays a part in the determination of the best performing data fusion method. From Table 5.5 and 5.6 it is clear that both the TCCRs and the model lifts are the lowest for data fusion method 'nearest neighbor'. From these two tables it is also clear that the data fusion method 'model based clustering approach' applied on the train data set has the highest TCCRs and model lifts. However, these statistics drop when applying the same data fusion model on the two test sets. This is the model overfitting phenomenon, as described above. This model overfitting can be seen in all the data fusion method used. However, it seems that for the data fusion method 'fusion value specific probability approach' this is the least. For the method 'fusion value specific probabilities approach' both the TCCRs and the model lifts are among the highest and the difference between the train data set and the test data sets is not as large as the other fusion methods.

Taking into account the three considerations, the fusion value specific proba-

Table 5.4: Frequencies after Applying Data Fusion Methods for Domain Energy

method		data set	#records	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
	actual	train	875	30.1%	19.8%	15.9%	17.7%	16.6%
		test1	411	33.8%	20.7%	15.6%	16.5%	13.4%
		test2	409	33.0%	19.3%	15.2%	20.0%	12.5%
nearest neighbor	predicted	train	875					
		test1	411	31.9%	19.7%	16.1%	18.0%	14.4%
		test2	409	32.2%	20.8%	17.4%	16.1%	13.4%
logistic regression	predicted	train	875	37.5%	21.5%	14.4%	13.4%	13.3%
		test1	411	39.4%	20.7%	15.6%	12.2%	12.2%
		test2	409	39.9%	17.1%	17.6%	16.4%	9.0%
fusion value specific approach	predicted	train	875	32.8%	22.7%	13.0%	12.3%	19.1%
		test1	411	39.2%	23.4%	13.4%	9.7%	14.4%
		test2	409	35.2%	22.7%	13.0%	14.2%	14.9%
model based clustering approach	predicted	train	875	33.1%	25.1%	13.3%	15.5%	12.9%
		test1	411	38.7%	25.5%	13.1%	11.4%	11.2%
		test2	409	35.5%	21.5%	12.7%	17.6%	12.7%

Table 5.5: Total Correct Classification Rates after Applying Data Fusion Methods for Domain Energy

method		data set	#records	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	total	chance
nearest neighbor	predicted	train	875							21.4%
		test1	411	40.5%	34.6%	24.2%	20.3%	22.0%	30.4%	22.7%
		test2	409	37.9%	25.9%	31.0%	25.8%	18.2%	29.6%	22.5%
logistic regression	predicted	train	875	50.9%	47.9%	48.4%	50.4%	50.9%	49.8%	21.4%
		test1	411	48.4%	40.0%	32.8%	24.0%	26.0%	38.7%	22.7%
		test2	409	46.6%	40.0%	33.3%	43.3%	32.4%	41.3%	22.5%
fusion value specific approach	predicted	train	875	47.7%	42.7%	46.5%	46.3%	43.1%	45.4%	21.4%
		test1	411	51.6%	44.8%	43.6%	37.5%	25.4%	43.8%	22.7%
		test2	409	54.2%	43.0%	49.1%	41.4%	34.4%	46.2%	22.5%
model based clustering approach	predicted	train	875	56.2%	51.0%	56.0%	50.7%	54.9%	54.1%	21.4%
		test1	411	45.9%	42.9%	38.9%	23.4%	23.9%	39.2%	22.7%
		test2	409	44.8%	37.5%	32.7%	29.2%	17.3%	35.5%	22.5%

Table 5.6: Model Lifts after Applying Data Fusion Methods for Domain Energy

method		data set	#records	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	total
nearest neighbor	predicted	train	875						
		test1	411	120	167	156	123	165	134
		test2	409	115	134	204	128	146	131
logistic regression	predicted	train	875	169	242	305	285	307	233
		test1	411	144	193	211	145	194	171
		test2	409	141	207	220	216	260	184
fusion value specific approach	predicted	train	875	159	216	293	261	260	213
		test1	411	152	217	280	227	190	193
		test2	409	164	223	324	206	276	205
model based clustering approach	predicted	train	875	187	262	353	286	331	253
		test1	411	136	207	250	141	179	173
		test2	409	136	194	216	146	139	158

Table 5.7: Frequency Clusters in Customer Database for Application Energy

Cluster	Frequency customers	Percentage customers
Cluster 1	334,083	29.5% (33.0%)
Cluster 2	204,774	18.1% (20.2%)
Cluster 3	165,416	14.6% (16.3%)
Cluster 4	176,319	15.6% (17.4%)
Cluster 5	131,970	11.6% (13.0%)
No cluster	120,843	10.7%
Total	1,133,405	100.0%

bilities approach turns out to be the best performing data fusion method. Consequently, this method is used to fuse the motivational clusters to the company's customer database. As a result of this data fusion the motivational clusters become 'known' for all the customers in the database. This is the starting point for differentiated marketing strategies as described in the next subsection.

External Validation

Despite the internal validation described in the previous subsection, the final validation is in the real world. Before a data fusion is proposed, the marketing company has a certain goal to achieve. This can be, for example, improving the response on a certain questionnaire, increasing sales, et cetera. External validation is done in order to draw conclusions about how this goal is achieved. It is clear that each marketing company has a different goal to achieve with data fusion, that's why there are no unified statistics for external validation. For each external validation tailor-made criteria need to be made.

Unfortunately external validation is not common practice for most marketing companies (Bell and Vincze, 1988, p.452). It is expensive and time-consuming. However, these type of cross-validation experiments are highly recommended. In the end it is not important what the statistics are in the internal validation step, but what the effect is of the differentiated marketing strategy in the real world. This is best described by the proverb '*the proof of the pudding is in the eating*'.

Keeping this proverb into consideration, the following external validation is performed for the application about the Dutch energy supplier: In the case of the supplier, the initial goal was to improve the response on the written questionnaire. From past research experiences the supplier knows that the response percentage on regular questionnaires is 19.9%. The first goal with the differentiated questionnaire

Table 5.8: Responses per Batch

Batch	Date sent	Questionnaires sent (#)	Response (#)	Response (%)
1	Oct 2002	549,818 (48.6%)	109,754 (38.5%)	20.0%
2	Oct 2002	260,151 (23.0%)	85,118 (29.8%)	32.7%
3	Nov 2002	217,928 (19.3%)	48,145 (16.9%)	22.1%
4	Nov 2002	103,508 (9.1%)	42,260 (14.8%)	40,8%
Total		1,133,405 (100.0%)	285,453 (100.0%)	25.2%

approach is to improve this response percentage.

The second goal is to improve the number of sales leads. The energy supplier defines the sales leads as the number of products or services, customers are interested in. In the questionnaire, customers are asked about their interests in ten energy products and services. From past experiences the supplier knows that the average number of sales leads is 2.25 per customer. The second goal with the differentiated questionnaire approach is to increase the average number of sales leads per customer.

As a results of the data fusion the total customer database with $1,133,405$ customers is classified. The columns 'Frequency customers' and 'Percentage customers' in Table 5.7 show the resulting motivational cluster frequencies of the fused data set $\widehat{\mathbf{Y}}$. For 120,843 (=10.7%) customers there are no or insufficient common items available in order to classify to one of the five motivational clusters.

Using the descriptions of the five motivational clusters, for each cluster a separate questionnaire is made by a specialized communication agency. For the group of customers with no motivational cluster, the regular questionnaire is used. The content of the questionnaires, that are the questions about the customer's interests in energy products and services, is the same for each questionnaire. Only the lay out (colors and pictures used in the questionnaire) and the tone-of-voice of the invitation letters are different for the cluster specific questionnaires. The focus of the questionnaire for motivational cluster 1 is on the balance between comfort and nature. The questionnaire for motivational cluster 2 emphasizes that the interests, wishes, desires, complaints, et cetera, from society, are taken seriously. For motivational cluster 3 the focus of the questionnaire is on the supplier's differentiated approach in order to increase the customer's comfort and to decrease the energy prices. The focus of the questionnaire for motivational cluster 4 is on, the saver the customer is with energy, the better it is for nature. And, finally, the focus of the questionnaire for motivational cluster 5 is on the question: 'Would you like to help us to improve our service for you?'.

Table 5.9: Responses per Motivational Cluster

Cluster	Questionnaires sent (#)	Response (#)	Response (%)
Cluster 1	334,083 (29.5%)	99,961 (35.0%)	29.9%
Cluster 2	204,774 (18.1%)	55,064 (19.3%)	26.9%
Cluster 3	165,416 (14.6%)	30,638 (10.7%)	18.5%
Cluster 4	176,319 (15.6%)	34,264 (12.8%)	19.4%
Cluster 5	131,970 (11.6%)	41,210 (14.4%)	31.2%
No cluster	120,843 (10.7%)	24,316 (8.5%)	20,1%
Total	1,133,405 (100.0%)	285,453 (100.0%)	25.2%

Eventually in four batches 1,133,405 (un)differentiated questionnaires were sent to all the customers. Table 5.8 shows when and how many questionnaires were sent to the customers in each batch. This table also shows how many customers responded on the questionnaires. Table 5.9 further splits these responses into the motivational clusters. From this table it is also clear that the first goal is attained. The total response percentage is 25.2%, which is higher than the target percentage of 19.9%. The difference in response percentage equals almost 60,000 extra customers, which is of course, valuable for the supplier.

Although the total response percentage in Table 5.9 displays 25.2% it is interesting to see what the response behavior is for each motivational cluster. From Table 5.9 it can be seen that the response percentages for the customers classified to motivational cluster 3 and 4, are relatively low. From past experiences with the motivational clusters it is known that customers classified to motivational cluster 3 and 4 are in general less willing to fill out questionnaires.

The second goal to attain, is increasing the number of sales leads. Table 5.10 shows the average number of sales leads per motivational cluster. From this table it is clear that also the second goal is attained; the average number of sales leads is 2.63, whereas an average of 2.25 sales leads was the target. Also from Table 5.10 it is interesting to see what the average number of sales leads is for each of the motivational clusters. The results in the table are completely consistent with the description of these five motivational cluster. Cluster 1 with a higher interest in energy products and services in order to get a good balance between own comfort and nature. Cluster 3 with a higher interest in energy products and services in order to get a differentiated approach for more comfort and lower prices. Cluster 5 with a higher interest in energy products, in order to stay in control with their own thoughts about energy. And, cluster 2 and 4 with a lower interest in energy products and

Table 5.10: Sales Leads per Motivational Cluster

Cluster	Response (#)	Sales leads
Cluster 1	99,961 (35.0%)	2.69
Cluster 2	55,064 (19.3%)	2.26
Cluster 3	30,638 (10.7%)	2.75
Cluster 4	34,264 (12.8%)	2.22
Cluster 5	41,210 (14.4%)	3.14
No cluster	24,316 (8.5%)	2.73
Total	285,453 (100.0%)	2.63

services, because they totally rely on the expertise of the energy supplier. However, there is no logical explanation for the fact that customers, not classified to one of the five motivational clusters, have a relative high average number of sales leads.

Although the responses and sales leads can be determined before and after the marketing strategy, it is impossible to conclude that the increase (or decrease) in responses and sales leads can be fully dedicated to the differentiated marketing strategy (Bell and Vincze, 1988, p.451; Kooiker, 1997, Section 8.4). When sending the questionnaires it was impossible to control for all kind of side effects that may be associated with response behavior and interests. However, for this applications both goals are attained: almost 60,000 customers more responded to the differentiated questionnaires and, on average, the total responding customers were more interested in energy products and services. Furthermore, instead of conducting a motivational research study among all 1,133,405 customers, only a small, representative number of these customers (1,751) where used. Which is, in terms of dollars, a huge saving in marketing research costs.

5.3 Discussion

In this chapter data sets were fused (or integrated) to each other. In order to be as realistic as possible this chapter used only real data sets. No simulated data sets were used, where inevitably, one could favor a simulation model and simulation parameters. The experiments described in this chapter were performed in their most realistic context.

In the marketing application the customer database of an energy supplier was fused to a motivational research study about energy. One of the most important goals was the reconstruction of customer's individual fusion values. Or, translated to the marketing application, it was undesirable that a customer, that belongs to motivational cluster 1, was fused to motivational cluster 2. In order to show the realistic number of such mismatches, this chapter concentrated on two very important validation steps, that were the internal validation step and the external validation step.

5.3.1 Internal Validation

The most important thing in the internal validation step was the prevention against model overfitting. The application showed that model overfitting was a serious problem. For example, in the case of the model based clustering approach the method showed the best statistics on the train data set, but subsequently failed to preserve these good statistics on the test data sets.

In order to prevent against model overfitting, this chapter used a train data set and two test data sets. The latter was done because of the dependency of the validation results of one particular split of the data set used. The train data set was used for training (or calibrating) the data fusion models and the two test data sets were used for validating the data fusion models.

The lesson that can be learnt from this is that, one should never trust a data fusion company that uses only one data set to train and test data fusion models. You have to take into account model overfitting, as we have shown using the train and test data sets.

In order to draw conclusions about the quality of the data fusion, this chapter used the statistics model lift and total correct classification rate (TCCR). The latter was calculated for both the random chance model and the data fusion model under study. In the application the fusion value specific probabilities approach was found to be the 'best' method. This is not only the case in the application described in this chapter, but also for past marketing applications in domains like care, insurance, gardening, financial services, et cetera (see track record on `www.smartagent.nl`). The problems and the goals of these marketing applications were similar to the application described in this chapter. In these past marketing applications the data fusion methods, as described in Section 5.2.3, were also used and compared. In each application the fusion value specific probabilities approach turned out to be (one of) the best methods in the internal validation, which makes this data fusion method, a method with stable results.

For the marketing application in this chapter the TCCR for the overall model was around 40%, whereas the TCCR with a random chance mode was around 20%. The model lift was around 200%, which means that the fusion value specific probabilities approach provided around 100% more correct matches than would be obtained by chance. Of course, the goal of the data fusion was to get a TCCR that was as close to 100% as possible, but when analyzing the TCCRs, we had to take into account the type of the fusion item and the type of the explanatory items. The fusion items in the application were motivational clusters that came from a motivational research study. The explanatory items were socio-demographical and socio-economical items. When it was possible to predict (almost) perfectly the motivational clusters with these explanatory items, the initial motivational research study would loose their uniqueness.

5.3.2 External Validation

As a result of the data fusion, the motivational clusters were estimated for all the customers in the database. In the real world application this was the starting point for differentiated marketing strategies. In the application about the energy supplier differentiated written questionnaires were made.

As *the proof of the pudding is in the eating*, the external validation step was even more important than the internal validation step. In the end the external validation step determined whether the data fusion was profitable or not.

Using a cross validation experiment, different marketing goals were tested and attained. In the case of the energy supplier almost 60.000 more customers responded on the differentiated questionnaires. Also the average number of sales leads per customer increased.

Given the large number of customers involved in the application, the increases in responses and sales leads gave the company a tremendous amount of extra information and sales opportunities. Furthermore, by using only a small proportion of the customers for a domain study, a lot of dollars were saved on marketing research costs. In the application the data fusion project was profitable and, consequently, was successful.

Chapter 6

Improving your Sales with Data Fusion

Abstract*

This chapter shows how an European mail order company uses data fusion in order to improve sales. To select the best data fusion algorithm, two traditional data fusion methods, that are polytomeous logistic regression and a nearest neighbor algorithm, are compared with two model based clustering approaches. Finally, it is shown how internal and external validation criteria are used in order to evaluate the results of the data fusion algorithms.

*This chapter is published as Van Hattum, P. and Hoijtink, H. (2009a). Improving your Sales with Data Fusion.*Journal of Database Marketing & Customer Strategy Management, 16(1)*, 7-14.

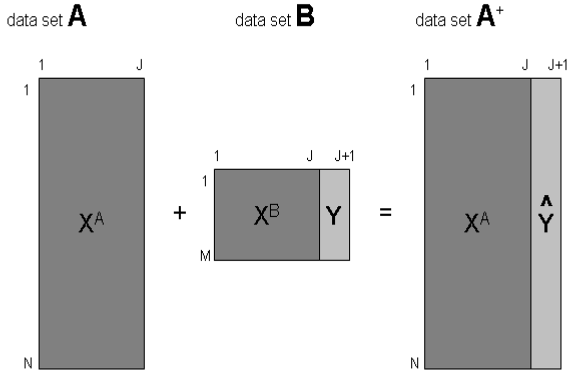


Figure 6.1: Schematic Representation of Data Fusion in Marketing (Derived and Adjusted from Van der Putten et al. (2002))

6.1 Introduction

In this chapter the following problem is addressed: an European mail order company specialized in gardening products, wants to send differentiated catalogues to all the customers in the database. In order to do this, information about customer gardening preferences and interests from an external database is fused to the customer database. Using the fused information, differentiated catalogues can be send to all the customers.

Above described problem can best be illustrated using the schematic representation in Figure 6.1. In this representation data set **A** is the customer database and contains knowledge and information (represented by J items) from all customers. External data set **B** contains knowledge and information (represented by $J + 1$ items) from a group of customers, that are not in data set **A**. The knowledge and information represented by the first J items is available for each customer in each data set. However, for the group of customers in external data set **B** there is some additional knowledge and information, that is item $J + 1$. The goal of this chapter is to fuse the extra knowledge and information in external data set **B**, that is item $J + 1$, to data set **A**. As a result of this data fusion the knowledge and information about item $J + 1$ becomes 'known' for all customers in the database, data set **A**.

The structure of this chapter is as follows. Section 6.2 describes how data fusion can be used in the context of the European mail order company. Also the marketing goals and the data sets used are described. Section 6.3 shows how data fusion is evaluated using internal and external validation criteria. This chapter concludes with a discussion in Section 6.4.

6.2 Application

6.2.1 Improving Sales

An European mail order company specialized in all kind of gardening products, that are flowers, bulbs, plants, et cetera, wants to increase the number of buying customers (Lattin and Bucklin, 1989; Van der Putten et al. 2002). In order to do this, they decide to develop a direct marketing strategy. From an external motivational research study, the mail order company knows that each individual has a different attitude towards gardening and gardening products (www.tuinbeleving.nl).

From the motivational study it is known that there are actually four groups (or clusters) of customers who have more of less the same attitude towards gardening and gardening products. Short descriptions (see www.tuinbeleving.nl for detailed descriptions) of these four motivational clusters are:

- cluster 1: gardens in this clusters are different from other gardens. They are surprising, wild, romantic and stylish; gardens meant for relaxation and unwinding. For customers in this cluster gardening brings that relaxation;

- cluster 2: gardens in this cluster are more often large patios, easy to maintain and cluttered. Gardens are outdoor spaces to hang out with family and friends. Customers in this cluster think of gardening as strenuous, rather than a relaxing activity;

- cluster 3: in this cluster the true gardener can be found. Gardens in this cluster are nice, neat, full of atmosphere and fit in with the rest of the neighborhood. Gardening is relaxing, a passion and the main hobby;

- cluster 4: gardens in this cluster are practical and easy to maintain. Customers in this cluster don't feel like gardening and can't find the time for gardening.

These motivational clusters provide a basis for developing a company's vision and/or a company's marketing directions on the strategic, tactical and operational levels, aligning the total marketing mix around the consumers needs in the domain gardening. Table 6.1 displays the frequencies of the resulting motivational clustering. Furthermore, each individual deals, handles and perceives gardening catalogues

Table 6.1: Frequency of Respondents in Motivational Clusters concerning Gardening (Between Brackets are the Percentages Based on the Total Number of Respondents Classified to One of the Four Motivational Clusters)

Cluster	Frequency respondents	Percentage respondents
Cluster 1	246	21.6% (23.0%)
Cluster 2	228	20.0% (21.3%)
Cluster 3	343	30.1% (32.0%)
Cluster 4	254	22.3% (23.7%)
No cluster	70	6.1%
Total	1,141	100.0%

in a different way. The mail order company assumes that giving customers (some sort of) a tailor made offer, will eventually increase the number of buying customers.

Using the descriptions of the four motivational clusters, for each of the four clusters a separate catalogue can be made by a specialized communication agency. The content of the catalogues, that is the gardening products offered to the customers, is the same for each catalogue. Only the lay out (colors and pictures used on the front page and the back page) and the tone-of-voice of the catalogue's introduction are different for the cluster specific catalogues.

Because the mail order company wants to send differentiated catalogues to their customers, data fusion is used. Using the common items in both the supplier's customer database and the external motivational research study, data fusion methods are used to fuse the motivational cluster to the supplier's customer database.

6.2.2 Description of the Data Set Used

The available data sets are the customer database of the company and the data set with the external motivational research study. Or, translated to Figure 6.1, data set **A** and data set **B**, respectively. The content of data set **A** is data set $\mathbf{X}^A$. This data set $\mathbf{X}^A$ contains $J = 7$ items, that are house ownership, number of vehicles, education, socio-demographic typology, household stage, prosperity and spending behavior, from all $N = 66,549$ customers.

The content of data set **B** is data set $\mathbf{X}^B$ and data set $\mathbf{Y}$. Data set $\mathbf{X}^B$ contains the same $J = 7$ items as data set $\mathbf{X}^A$. The content of data set $\mathbf{Y}$ is one item, that are the motivational clusters. In total $M = 1,141$ respondents has participated to the motivational study and for them the motivational clusters are known. Note that these $1,141$ customers are not a fraction of the $66,549$ customers from data set $\mathbf{A}$.

As described in Section 6.1 and illustrated in Figure 6.1, the goal of the data fusion process is to fuse the information in data set $\mathbf{Y}$ to data set $\mathbf{A}$ using the common items in $\mathbf{X}^A$. Or, in the context of this application, using data fusion methods, all the $66,549$ customers are classified to one of the four motivational clusters using the 7 common items. How good or bad this data fusion process is done, is described in the following subsections.

6.3 Validation

6.3.1 Internal Validation

Van Hattum and Hoijtink (2008) compare four data fusion methods, that are polytomeous logistic regression, a nearest neighbor algorithm, a fusion value specific probabilities method and a model based clustering approach. In the nearest neighbor algorithm the missing motivational clusters in data set $\mathbf{A}$ are duplicated from data set $\mathbf{B}$ using cases with similar values on the explanatory items. In polytomeous logistic regression the relationship between motivational cluster and the explanatory items, as determined in data set $\mathbf{B}$, is used to fuse information to data set $\mathbf{A}$. The fusion value specific probabilities model is based on latent cluster analysis, where the role of the latent clusters is taken by the fusion value, in this case the motivational cluster and the explanatory variables are the items. The model fitted in data set $\mathbf{B}$ is used to predict the motivational clusters in data set $\mathbf{A}$. The assumption in the model based clustering approach is that the data are generated from a mixture of fusion values specific probabilities models. As a result of the model based clustering approach there will be a fusion value specific probabilities model for each latent cluster found. Translated to the application at hand, the number of latent clusters found is 9; for each of the 9 latent clusters a fusion value specific probabilities model is estimated. The interested reader is referred to Van Hattum and Hoijtink (2008) for a full description of the four data fusion approaches.

In order to select the best data fusion method, the statistics TCCR (Total Correct Classification Rate) and model lift are calculated. The TCCR is the percentage of respondents that are classified to the right motivational cluster. Furthermore, a percentage of correct classifications based on a random chance model can be obtained. Or, in other words, the percentage of correct classifications that can be expected when the motivational clusters are randomly assigned to the respondents. This percentage is called TCCR$_{chance}$. The statistic model lift is calculated as $\frac{TCCR}{TCCR_{chance}} * 100\%$ and can be interpreted as the percentage of more correct classifications than would be obtained by chance. All statistics can be calculated for the overall model and for each motivational cluster separately.

Above described statistics are calculated for each data fusion method applied to one training data set (about 50% of the cases) and two test data sets (each about 25% if the cases). Models fitted on a data set tend to predict much better for that data set than for a new data set sampled from the same population. Since we want to fuse values from data set B to data set A, we will evaluate the predictive performance of the four fusion methods by fitting them on the training data set and using the resulting models to make predictions in the two test data sets. As will be illustrated below, this prevents against model overfitting. The interested reader is referred to Verstraeten (2005) for a further discussion of model overfitting.

Frequencies, TCCRs and model lifts can be used to determine which data fusion method performs the best. There must be a good comparison between the actual and the predicted frequencies of the motivational clusters. These frequencies are displayed in Table 6.2. Looking at this table, it is not clear which data fusion method to choose. When looking at, for example, the data fusion method 'logistic regression', it can be seen that the predicted frequency for motivational cluster 1 can be compared with the actual frequency. However, the difference between the predicted frequency for motivational cluster 2 and the actual frequency for this motivational cluster, is quite large. Similar observations can be made for other data fusion methods.

As can be seen from the TCCRs (displayed in Table 6.3) and the model lifts (displayed in Table 6.4) the data fusion method 'nearest neighbor' performs the worst of all. Both the TCCRs and the model lifts are the lowest compared to the other methods. The method 'model based clustering approach (with 9 latent clusters)' has the highest statistics applied to the train data set, but, due to model overfitting, these statistics drop, when applied to the test data sets. The other data fusion methods also suffer from model overfitting, but not as dramatic as the model based clustering approach. From the two tables it can be seen that the statistics for the data fusion methods 'logistic regression' and 'fusion value specific probabilities approach' are among the highest. Both methods are good models, however, the statistics for the latter model are more consistent on the train data set and the two test data sets.

Like in Van Hattum and Hoijtink (2008) the fusion value specific probabilities approach turns out to be the best performing data fusion method. Consequently, this method is used to fuse the motivational clusters to the company's customer database. As a result of this data fusion the motivational clusters become known for all the customers in the database. This is the starting point for differentiated marketing strategies as described in the next subsection.

Table 6.2: Classification Percentages after Applying Data Fusion Methods for Domain Gardening

method		data set	#records	cluster 1	cluster 2	cluster 3	cluster 4
	actual	train	539	23.0%	18.2%	34.3%	24.5%
		test1	257	21.4%	22.2%	30.4%	26.1%
		test2	275	24.4%	26.5%	29.1%	20.0%
nearest neighbor	predicted	train	539				
		test1	257	25.7%	17.1%	36.2%	21.0%
		test2	275	25.5%	18.5%	33.5%	22.5%
logistic regression	predicted	train	539	24.7%	10.9%	38.6%	25.8%
		test1	257	23.3%	11.3%	38.9%	26.5%
		test2	275	25.8%	12.7%	39.3%	22.2%
fusion value specific approach	predicted	train	539	33.2%	13.7%	31.7%	21.3%
		test1	257	31.1%	16.3%	29.6%	23.0%
		test2	275	32.0%	15.3%	34.5%	18.2%
model based clustering approach	predicted	train	539	25.4%	13.2%	39.3%	22.1%
		test1	257	23.3%	16.0%	34.2%	16.5%
		test2	275	21.8%	17.8%	35.3%	25.1%

Table 6.3: Total Correct Classification Rates after Applying Data Fusion Methods for Domain Gardening

method		data set	#records	cluster 1	cluster 2	cluster 3	cluster 4	total	chance
nearest neighbor	predicted	train	539						26.4%
		test1	257	31.8%	20.5%	38.7%	37.0%	33.5%	25.5%
		test2	275	38.6%	35.3%	37.0%	22.6%	33.8%	25.4%
logistic regression	predicted	train	539	46.6%	50.8%	55.8%	51.1%	51.8%	26.4%
		test1	257	43.3%	48.3%	43.0%	52.9%	46.3%	25.5%
		test2	275	42.3%	57.1%	40.7%	37.7%	42.5%	25.4%
fusion value specific approach	predicted	train	539	39.1%	45.9%	55.0%	47.0%	46.8%	26.4%
		test1	257	41.2%	45.2%	44.7%	55.9%	46.3%	25.5%
		test2	275	44.3%	52.4%	46.3%	36.0%	44.7%	25.4%
model based clustering approach	predicted	train	539	48.9%	47.7%	58.5%	52.9%	54.7%	26.4%
		test1	257	40.0%	31.7%	42.0%	44.1%	40.5%	25.5%
		test2	275	43.3%	42.9%	39.2%	31.9%	38.9%	25.4%

Table 6.4: Model Lifts after Applying Data Fusion Methods for Domain Gardening

method		data set	#records	cluster 1	cluster 2	cluster 3	cluster 4	total
nearest neighbor	predicted	train	539					
		test1	257	149	92	128	142	131
		test2	275	158	133	127	113	133
logistic regression	predicted	train	539	203	280	162	209	196
		test1	257	202	218	142	203	182
		test2	275	173	215	140	189	167
fusion value specific approach	predicted	train	539	170	253	160	192	177
		test1	257	193	204	147	215	182
		test2	275	182	197	159	180	176
model based clustering approach	predicted	train	539	213	318	170	216	208
		test1	257	187	143	139	169	159
		test2	275	178	161	135	159	153

Table 6.5: Frequency Clusters in Application Gardening (Between Brackets are the Percentages Based on the Total Number of Customers Classified to One of the Four Motivational Clusters)

Cluster	Frequency customers	Percentage customers
Cluster 1	16,938	25.5% (28.0%)
Cluster 2	8,409	12.6% (13.9%)
Cluster 3	24,076	36.2% (39.8%)
Cluster 4	11,070	16.6% (18.3%)
No cluster	6,056	9.1%
Total	66,549	100.0%

6.3.2 External Validation

In the case of the European mail order company the initial goal was to increase the number of buying customers. From past experiences the mail order company knows that 3.58% of the customers, who received a catalogue, bought something from this catalogue within four weeks after receival. Using differentiated catalogues, the goal of the mail order company is to increase this number of buying customers.

As a results of the data fusion the total customer database, with $66,549$ customers is classified. The columns 'Frequency customers' and 'Percentage customers' in Table 6.5 show the resulting motivational cluster frequencies of the fused data set $\widehat{\mathbf{Y}}$. For 6,056 (=9.1%) customers there are no or insufficient common items available in order to classify to one of the four motivational clusters. From the percentages between brackets in Table 6.1 and 6.5 it is clear that the mail order company has more customers classified to motivational cluster 1 and 3, compared with the external motivational research study. This is completely consistent with the description of these two motivational clusters; both motivational clusters contain in general more true and passionate gardeners.

Using the descriptions of the four motivational clusters, for each cluster a separate catalogue can be made. However, in the first step of the differentiated marketing strategy, the mail order company wants to concentrate on just two of the four motivational clusters, that are motivational clusters 1 and 3. So, for only these two motivational clusters, cluster specific catalogues are made by a specialized communication agency. The content of the catalogues, that is the gardening products offered to the customers, is the same. Only the lay out (colors and pictures used on the front page and the back page) and the tone-of-voice of the catalogue's introduction are different for the cluster specific catalogues. The focus of the catalogue for

Table 6.6: Number of Catalogues Sent

		Catalogue sent	
		Standard	cluster specific
Predicted	cluster 1	6,250	6,250
	cluster 3	6,250	6,250

motivational cluster 1 is on getting inspired by the catalogue, self creation of gardens, exotic and adventurous gardens. The focus of the catalogue for motivational cluster 3 is on traditional and hobby gardening, and on the amount of information that is giving about gardening products and services.

Furthermore, in order to validate the two cluster specific catalogues, the mail order company decides to compare the results with the standard catalogue. This is done using a randomized experiment among a sample from the customers who are classified to either motivational cluster 1 or 3. The set-up for the randomized experiment is as follows (see also Table 6.6):

- To 6250 customers, who are classified to motivational cluster 1, standard catalogues are sent;
- To 6250 customers, who are classified to motivational cluster 1, cluster 1 specific catalogues are sent;
- To 6250 customers, who are classified to motivational cluster 3, standard catalogues are sent;
- To 6250 customers, who are classified to motivational cluster 3, cluster 3 specific catalogues are sent.

From the 25,000 customers, who are in the randomized experiment, the percentages of customers, as displayed in Table 6.7, bought something from the (un)differentiated catalogues. From this table it is clear that the differentiated catalogue approach rendered more buying customers than with the standard catalogues. Not only compared with the customers who received undifferentiated catalogues, but also with the 3.58% buying customers from past experiences. It can be concluded that the company's goal with the differentiated marketing strategy is attained.

Also from Table 6.7 it is interesting to see the difference between the two motivational clusters. From the table it is clear that the effect of the differentiated catalogue is larger for motivational cluster 1 than for motivational cluster 3. This difference can be explained by the following three arguments. First of all, from

Table 6.7: Percentage Buying Customers

		Catalogue sent	
		Standard	Cluster specific
Predicted	cluster 1	3.70%	4.46%
	cluster 3	3.65%	3.83%

past experiences with the motivational clusters it is known that customers classified to motivational cluster 1 are in general more sensitive to (perceived) tailor made offerings than customers classified to motivational cluster 3. Secondly, from past experiences with the motivational clusters it is known that customers classified to motivational cluster 1 buy in general more products from mail order companies. And finally, the specialized communication agency thinks that the lay out of the standard catalogue and the tone-of-voice of the standard catalogue's introduction are more likely to attract customers classified to motivational cluster 3.

However, for both motivational clusters 1 and 3 the percentages of buying customers is higher using a differentiated marketing approach than with a undifferentiated or standard approach. With the differentiated catalogue the European mail order company is able to get more buying customers and, consequently, increase his turnover.

The researcher must keep in mind, that it is impossible to conclude that the increase in number of buying customers can be fully dedicated to the differentiated catalogues. When sending the catalogues it was impossible to control for all kind of side effects that may be associated with customer buying behavior. However, for this application the goal of increasing the number of buying customers is attained. Furthermore, by using the results of an external motivational research study instead of conducting there own research study, the mail order company saved dollars on marketing research activities.

6.4 Discussion

In this chapter, the customer database of a mail order company was fused to a motivational research study about gardening. In order to fuse the data sets, different traditional and new data fusion methods were used in order to fuse the data sets.

In the internal validation step the different data fusion methods were compared and the fusion value specific probabilities approach was found to be the 'best' method. The TCCR for the overall model was around 45-46%, whereas the TCCR with a random chance mode was around 25%. The resulting model lift was around 180%, which means that the fusion value specific probabilities approach provided around 80% more correct matches than would be obtained by chance. The conclusion that the fusion value specific probabilities approach performed the 'best' was not only drawn in this chapter, but also in past research (Van Hattum and Hoijtink, 2008). This makes this data fusion method, a method with stable results.

As a result of the internal validation step the motivational clusters were estimated for all the customers in the database. This was the starting point for differentiated marketing strategies, or, in the case of the mail order company, differentiated catalogues were made by a specialized company.

Using a randomized experiment different marketing goals were tested. In the case of the mail order company more customers bought something from the catalogue, when receiving the right (differentiated) catalogues.

Given the large number of customers involved, the increase in buying behavior gave the company a tremendous amount of extra sales. Furthermore, by using an external domain study, a lot of dollars were saved on marketing research costs. In all cases the data fusion project was profitable, and consequently, was successful.

References

Agresti, A. (2002). *Categorical Data Analysis.* New York: John Wiley.

Ainslie, A. and Rossi, P.E. (1998). Similarities in Choice Behavior across Product Categories. *Marketing Science, 17(2)*, 91-106.

Akaike, H. (1987). Factor Analysis and AIC. *Psychometrika, 52(3)*, 317-332.

Alderson, W. (1965). *Dynamic Marketing Behavior.* Illinois: Homewood.

Andrews, R.L. and Currim, I.S. (2003). A Comparison of Segment Retention Criteria for Finite Mixture logit models. *Journal of Marketing Research, 40(2)*, 235-243.

Baker, K., Harris, P. and O'Brien, J. (1989). Data Fusion: An Appraisal and Experimental Evaluation. *Journal of the Market Research Society, 31(2)*, 153-212.

Banfield, J.D. and Raftery, A.E. (1993). Model Based Gaussian and Non-Gaussian Clustering. *Biometrics, 49(3)*, 803-821.

Bell, M.L. and Vincze, J.W. (1988). *Managerial Marketing: Strategy and Cases.* New York: Elsevier Science Publishing Co..

Bensmail, H., Celeux, G., Raftery, A.E. and Robert, C.P. (1997). Inference in Model Based Clustering. *Statistics and Computing, 7(1)*, 1-10.

Bozdogan, H. (1987). Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions. *Psychometrika, 52(3)*, 345-370.

Brethouwer, W., Lamme, A., Rodenburg, J., Du Chatinier, H. and Smit, M. (1995). *Quality Planning toegepast (Dutch).* Amsterdam: Janssen Offset.

Bronner, A.E. (1988). Einde fusie-fobie in Nederland? *Jaarboek van de Nederlandse vereniging van Marktonderzoekers, 1988*, 9-18.

Buck, S. (1989). Single Source Data-The Theory and the Practice. *Journal of the Market Research Society, 31(4)*, 489-500.

Bucklin, R.E. and Gupta, S. (1992). Brand Choice, Purchase Incidence and Segmentation: An Integrated Modelling Approach. *Journal of Marketing Research, 29(2)*, 201-215.

Buckinx, W. (2005). Using Predictive Modeling for Targeted Marketing in a Non-Contractual Retail Setting. PhD. thesis, Marketing, Gent University, Belgium.

Bull, N.H. and Passewitz, G.R. (1994). Finding Customers: Market Segmentation. *Publication CDFS-1253-94*. Ohio State University.

Callebaut, J., Janssens, M., Op de Beeck, D., Lorré, D. and Hendrickx, H. (1999). *Motivational Marketing Research Revisited*. Leuven: Garant Publishers.

Congdon, P. (2005). *Bayesian Models for Categorical Data*. New York: John Wiley.

Cowles, M.K. and Carlin, B.P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association, 91*, 833-904.

Craig, C.S. and McCann,J.M. (1978). Item Nonresponse in Mail surveys: Extent and Correlates. *Journal of Marketing Research, 15(2)*, 285-289.

Cui, G. and Choudhury, P. (2002). Marketplace Diversity and Cost-Effective Marketing Strategies. *Journal of Consumer Marketing, 19(1)*, 54-73.

Cui, G. and Choudhury, P. (2003). Consumer Interests and the Ethical Implications of Marketing: A Contingency Framework. *Journal of Consumer Affairs, 37(2)*, 364-387.

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Chichester: John Wiley & Sons Ltd.

Dempster,A.P., Laird,N.M. and Rubin,D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B, 39(1)*, 1-38.

Dillon, W.R., Goldstein, M. and Schiffman, L.G. (1978). Appropriateness of Linear Discriminant and Multinomial Classification Analysis in Marketing Research. *Journal of Marketing Research, 15(1)*, 103-112.

DiStefano, C. and Kamphaus, R. W. (2006). Investigating Subtypes of Child Development A Comparison of Cluster Analysis and Latent Class Cluster Analysis in Typology, *Educational and Psychological Measurement, 66(5)*, 778-794.

Elmore-Yalch, R. (1998). A Handbook: Using Market Segmentation to Increase Transit Ridership. *Report 36.*

Everitt, B.S. (1988). A Monte Carlo Investigation of the Likelihood Ratio Test for Number of Classes in Latent Class Analysis. *Multivariate Behavioral Research, 23(4)*, 531-538.

Feinberg, F.F., Krishna, A. and Zhang, J.Z. (2002). Do We Care What Others Get? A Behaviorist Approach to Targeted Promotions. *Journal of Marketing Research, 39(3)*, 277291.

Fraley, C. and Raftery, A.E. (1998). How Many Clusters? Which Clustering Method? Answers via Model Based Cluster Analysis. *Technical Report No. 329*, Department of Statistics, University of Washington.

Frank, R.E. (1972). Predicting New Product Segments. *Journal of Advertising Research, 12(3)*, 9-13.

Gelman, A., Meng, X. and Stern, H.S. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies (with discussion). *Statistica Sinica, 6(4)*, 733-807.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2000). *Bayesian Data Analysis*. London: Chapman and Hall.

Gilula, Z., McCulloch, R.E and Rossi, P.E. (2006). A Direct Approach to Data Fusion. *Journal of Marketing Research, 43(1)*, 73-83.

Haberman, S.J. (1988). A Stabilized Newton-Raphson Algorithm for Log-Linear Models for Frequency Tables derived by Indirect Observations. *Sociological Methodology, 1988*, 193-221.

Hagenaars, J.A. (1988). Latent Structure Models with Direct Effects between Indicators: Local Dependence Models. *Sociological Methods and Research, 16(3)*, 379-404.

Hagenaars, J.A. and Mccutheon, A.L. (2002). *Applied Latent Class Analysis.* Cambridge: Cambridge University Press.

Hair, J.F., Anderson, R.E., Tatham, R.L. and Black, W.C. (1984). *Multivariate Data Analysis.* London: Prentice-Hall International, Inc.

Hoijtink, H. (1998). Constrained Latent Class Analysis using the Gibbs Sampler and Posterior Predictive P-values: Applications to Educational Testing. *Statistica Sinica, 8*, 691-711.

Hoijtink, H. (2000). Posterior Inference in the Random Intercept Model Based on Samples obtained with Markov Chain Monte Carlo Methods. *Computational Statistics, 15(3)*, 315-336.

Hoijtink, H. (2001). Confirmatory Latent Class Analysis: Model Selection using Bayes Factors and (Pseudo) Likelihood Ratio Statistics. *Multivariate Behavioral Research, 36(4)*, 563-588.

Hoijtink, H. and Notenboom, A. (2004). Model Based Clustering of Large Data Sets: Tracing the Development of Spelling Ability. *Psychometrika, 69(3)*, 481-498.

Heidegger, M. (1991). Over denken, bouwen, wonen (Dutch). *Four essays*. Translated by H.M. Berghs.

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression*. Hoboken: John Wiley & Sons.

Jones, J.M. and Zufryden, F.S. (1980). Adding Explanatory Variables to Consumer Purchase Behavior Model: An Exploratory Study. *Journal of Marketing Research, 17(3)*, 323-334.

Kamakura, W.A. and Wedel, M. (1997). Statistical Data Fusion for Cross-Tabulation. *Journal of Marketing Research, 34(4)*, 485-498.

Kamakura, W.A., Wedel, M., De Rosa, F. Mazzon, J.A. (2003). Cross-Selling through Database Marketing: a Mixed Data Factor Analyzer for Data Augmentation and Prediction. *International Journal of Research in Marketing, 20(1)*, 45-65.

Kooiker, R. (1997). *Marktonderzoek. (Dutch)*. Groningen: Wolters-Noordhoff bv.

Lattin, J.M. and Bucklin, R.E. (1989). Reference Effects of Price and Promotion on Brand Choice Behavior. *Journal of Marketing Research, 26(3)*, 299-310.

Lin, T.H. and Dayton, C.M. (1997). Model Selection Information Criteria for Non-Nested Latent Class Models. *Journal of Educational and Behavioral Statistics, 22(3)*, 249-264.

Little, R.J. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley.

Maclachlan, D.L. and Mulhern, M. (2004). Segment Optimization. An Empirical Comparison. *ESOMAR 2004 Conference Proceedings ESOMAR: Amsterdam, The Netherlands*, 289-308.

Magidson, J. and Vermunt, J.K. (2002). Latent Class Models for Clustering: A Comparison with K-means. *Canadian Journal of Marketing Research, 20(1)*, 36-43.

Manchanda, P., Ansari, A. and Gupta, S. (1999). The "Shopping Basket": A model for Multicategory Purchase Incidence Decisions. *Marketing Science, 18(2)*, 95-114.

Meila, M. and Heckerman, D. (2001). An Experimental Comparison of Model Based Clustering Methods, *Machine Learning, 42(1)*, 9-29.

Mela, C.F., Gupta, S. and Lehmann, D.R. (1997). The Long-term Impact of Promotion and Advertising on Consumer Brand Choice. *Journal of Marketing Research, 34(2)*, 248-261.

Meng, X.L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22(3)*, 1142-1160.

Moustaki, I. and Papageorgiou, I. (2005). Latent Class Models for Mixed Variables with Applications in Archaeometry. *Computational Statistics and Data Analysis, 48(3)*, 659-675.

Mulhern, M. and MacLachlan, D.L. (2003). Using Latent Class Models to Improve Marketing Decisions: A Segmentation Illustration. *Canadian Journal of Marketing Research, 21*, 25-30.

Narayanan, A. (1990). Computer Generation of Dirichlet Random Vectors. *Journal of Statistical Computation and Simulation, 36(1)*, 19-30.

Newcomb, S. (1886). A Generalized Theory of the Combination of Observations So As To Obtain the Best Result. *American Journal of Mathematics, 8(4)*, 343-366.

Oppenhuisen, J. (2000). *Een schaap in de bus? Een onderzoek naar waarden van de Nederlander (Dutch).* Amsterdam: Grafische Producties.

Pearson, K. (1894). Contributions to the Mathematical Theory of Evolution. *Philosophical Transactions, A, 185*, 71-110.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches.* Lecture Notes in Statistics, 168, New York: Springer.

Ratner, B. (2003). *Statistical Modelling and Analysis for Database Marketing: Effective Techniques for Mining Big Data.* Florida: Chapman&Hall/CRC.

Reboussin, B.A., Ip, E.H. and Wolfson, M. (2008). Locally Dependent Latent Class Models with Covariates: An Application to Under-Age Drinking in the USA. *Journal of the Royal Statistical Society, Series A, 171(4)*, 877-897.

Richardson, S. and Green, P.J. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society, Series B, 59(4)*, 731-792.

Rodgers, W.L. (1984). An Evaluation of Statistical Matching. *Journal of Business and Economic Statistics, 2(1)*, 91-105.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* London: Chapman and Hall.

Schafer, J.L. and Graham, J.W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7(2)*, 147-177.

Smith, W. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Marketing, 21(1)*, 3-8.

Stanton, J. and Pires, G.D. (1999). The Substantiality Test: Meaning and Application. *Journal of Segmentation in Marketing, 3(2)*, 105-115.

Stevens, M. (2000). Dealing with Label Switching in Mixture Models. *Journal of the Royal Statistical Society, Series B, 62*, 795-810.

Ter Braak, C.J.F., Hoijtink, H., Akkermans, W. and Verdonschot, P.F.M. (2003). Bayesian Model Based Cluster Analysis for Predicting Macrofaunal Communities. *Ecological Modeling, 160(3)*, 235-248.

Van der Putten, P., Kok, J.N. and Gupta, A. (2002). Data Fusion Through Statistical Matching. *Paper 185.* MIT Sloan School of Management.

Van Hattum, P. and Hoijtink, H. (2008). The Proof of the Pudding is in the Eating. Data fusion: An Application in Marketing.*Journal of Database Marketing & Customer Strategy Management, 15(4)*, 267-284.

Van Hattum, P. and Hoijtink, H. (2009a). Improving your Sales with Data Fusion. *Journal of Database Marketing & Customer Strategy Management, 16(1)*, 7-14.

Van Hattum, P. and Hoijtink H. (2009b). Market Segmentation using Brand Strategy Research: Bayesian Inference with respect to Mixtures of Log-Linear Models. *Journal of Classification, 16.*

Varki, S. and Chintagunta, P.K. (2004). The Augmented Latent Class Model: Incorporating Additional Heterogeneity in Latent Class Model for Panel Data. *Journal of Marketing Research, 41(2)*, 226-233.

Verhage, B. and Cunningham, W.H. (1984). *Grondslagen van het Marketing Management (Dutch).* Leiden: H.E. Stenfert Kroese B.V.

Vermunt, J.K. (1997). *LEM 1.0: A General Program for the Analysis of Categorical Data.* Tilburg: Tilburg University.

Vermunt, J.K. and Magidson J. (2000). *Latent Gold.* Belmont: Statistical Innovations Inc.

Verstraeten, G. (2005). Issues in Predictive Modelling of Individual Customer Behavior: Applications in Targeted Marketing and Consumer Credit Scoring. PhD. thesis, Marketing, Gent University, Belgium.

Wang H., Obremski T., Alidaee, B. and Kochenberger, G. (2008). Clique Partitioning for Clustering: A Comparison with K-Means and Latent Class Analysis, *Communications in Statistics - Simulation and Computation, 37(1)*, 1 - 13.

Wedel, M. (1997). *GLIMMIX Users Manual*, Groningen: ProGAMMA.

Wedel, M. and Kamakura, W.A. (2000). *Market Segmentation: Conceptual and Methodological Foundations*. Norwell: Kluwer Academic Publishers.

Wind, J. and Mahajan, V. (1997). Editorial: Issues and Opportunities in New Product Development: An Introduction to the Special Issue. *Journal of Marketing Research, 34(1)*, 1-12.

Zeger, S.L. and Karim, M.R. (1991). Generalized Linear Models with Random Effects: A Gibbs Sampling Approach. *Journal of the American Statistical Association, 86*, 79-86.

Samenvatting

Dit proefschrift behandelt twee basisproblemen in de marketing, namelijk marktsegmentatie, dat is het groeperen van mensen met dezelfde eigenschappen, en marktbenadering, dat is het focussen van de marketinginspanningen op de meest aantrekkelijke marktsegmenten.

Om marktsegmentatie te conceptualiseren, stel je dan eens voor dat je in een heteluchtballon zit. Als je naar de personen beneden je kijkt, lijken ze erg op elkaar. Naarmate de ballon daalt, worden steeds meer verschillen zichtbaar; je ziet kleine en grote personen, slanke en dikke. Als de ballon verder daalt naar straatniveau en je begeeft je onder de personen op de straat, zul je ontdekken dat elke persoon iets unieks heeft, maar dat er ook veel overeenkomsten zijn die je niet vanuit de hoogte kon zien. Je ziet dat er mannen en vrouwen zijn, personen die netjes gekleed zijn en anderen die meer casual gekleed zijn. Sommige personen zijn zichtbaar vrolijk, anderen weer niet. Als je de personen aanspreekt, zul je nog veel meer overeenkomsten aantreffen. Sommige personen hebben een avontuurlijke houding in het leven, anderen wat meer ingetogen. Sommige personen houden ervan om in een luxe auto te rijden, waar anderen helemaal niks geven om auto's. Het lijkt erop dat deze personen tegelijkertijd op elkaar lijken, maar toch ook verschillend zijn. Deze tocht met de heteluchtballon weerspiegelt het concept van marktsegmentatie. Allereerst bekijk je de groep personen in zijn geheel. Wanneer je de groep nader bekijkt, zie je karakteristieken waarin de personen uit de groep kunnen verschillen. Als laatste kijk je naar de verschillende groepen van personen met overeenkomstige karakteristieken.

De meeste statistici hebben niet het voorrecht om een segmentatiestudie te doen vanuit een heteluchtballon en zijn aangewezen op databestanden. Deze databestanden zijn een versimpeling van de werkelijkheid en bevatten karakteristieken, vaak verzameld door middel van marktonderzoek, van de te onderzoeken populatie. Het doel in deze databestanden is het zoeken van groepen mensen met dezelfde eigenschappen. Om een simpel voorbeeld te geven: wanneer we aan een groep respondenten (dit zijn personen die mee hebben gedaan aan een marktonderzoek) vragen om tenminste een item te kiezen uit de volgende lijst van statements:

1. het is belangrijk om het nieuwste automodel te hebben,

2. het is belangrijk om zoveel mogelijk accessoires op mijn auto te hebben,

3. de prijs van de auto moet zo goedkoop mogelijk zijn,

4. de reden om een auto te hebben, is dat ik niet nat wordt,

zullen we waarschijnlijk twee marktsegmenten of clusters van personen vinden met min of meer dezelfde houding ten aanzien van auto's. Cluster 1 met personen die een hogere kans hebben om statement 1 en 2 te kiezen en een lagere kans hebben om statement 3 en 4 te kiezen. Voor personen uit dit cluster is een luxe auto een echte 'must have'. En Cluster 2 met personen die een hogere kans hebben om statement 3 en 4 te kiezen en een lagere kans hebben om statement 1 en 2 te kiezen. Voor personen uit dit cluster is een auto maar een middel om je van A naar B te brengen.

Voor dit simpele voorbeeld is het duidelijk wat voor soort clusters er kunnen worden verwacht. Echter, in marketing bevatten de databestanden vaak veel items en veel respondenten. In grote databestanden is het niet duidelijk hoeveel clusters er zijn en hoe deze clusters kunnen worden geïdentificeerd. Om de clusters in deze grote databestanden te vinden, zijn statistische clustertechnieken nodig. In feite, de meeste literatuur over marktsegmentatie gaat over de technieken om clusters te identificeren in databestanden. Een groot deel van deze literatuur zijn vergelijkende artikelen die de meest gebruikte clustertechnieken contrasteren (MacLachlan and Mulhern, 2004). Meer recente artikelen (MacLachlan and Mulhern, 2004; Magidson and Vermunt, 2002; Mulhern and MacLachlan, 2003) vergelijken modelgebaseerde clustertechnieken met meer traditionele clustertechnieken, zoals K-means.

In de context van segmentatie geven sommige artikelen aan betere segmentatieresultaten te vinden, wanneer gebruik wordt gemaakt van modelgebaseerde clustertechnieken (MacLachlan and Mulhern, 2004). Een belangrijk voordeel van modelgebaseerde clustering (Bensmail et al.,1997; Fraley and Raftery, 1998; Vermunt and Magidson, 2000, p. 1-2, 152) ten opzichte van traditionele clustertechnieken (Hair et al., 1984, p. 469-518) is het statistisch raamwerk waarop modelgebaseerde clustering is gebaseerd. Een nadeel van modelgebaseerde clusterbenaderingen ten opzichte van traditionele clustertechnieken is dat ze minder beschikbaar zijn in populaire statistische software. Dit resulteert in onderzoekers die hun eigen software maken, zoals bijvoorbeeld: Glimmix (Wedel and Kamakura, 2000, p.181-186) en LatentGold (Vermunt and Magidson, 2000).

In hoofdstuk 2, *'Market Segmentation using Brand Strategy Research: Bayesian Inference with respect to Mixtures of Log-Linear Models'*, gepubliceerd als Van Hattum and Hoijtink (2009b), wordt een Bayesiaanse modelgebaseerde clusterbenadering voor dichotome items gepresenteerd. Dit hoofdstuk laat zien hoe de clusterbenadering omgaat met missende waarden, grote databestanden en binnen cluster item afhankelijkheden. Verder worden de consequenties weergegeven als er een cluster model, waarin wordt aangenomen dat de items lokaal onafhankelijk zijn binnen de clusters, wordt gebruikt, terwijl in de data de items lokaal afhankelijk zijn binnen de clusters. De voorbeelden in dit hoofdstuk worden geïllustreerd met behulp van Brand Strategy Reseach, dat is een theoretisch raamwerk om motivationele groepen of clusters te maken.

Ondanks dat de clusterbenadering in Hoofdstuk 2 het statistisch optimale aantal clusters geeft, bevat deze oplossing vaak (vooral in het geval van grote databestanden) teveel clusters voor de beoogde marketingdoeleinden. Hieruit wordt duidelijk dat marktsegmentatie niet alleen een kwestie is van statistiek, maar een interactie tussen statistiek en marketing. Of, zoals MacLachlan en Mulhern (2004) dit goed verwoorden: *in elk empirisch probleem moet de onderzoeker noodzakelijkerwijs een behoorlijke dosis subjectiviteit en domeinkennis gebruiken. Dit kan door het berekenen van een paar statistische indicatoren, maar uiteindelijk wordt de beslissing over het aantal clusters genomen door deze indicatoren te interpreteren in het licht van het marketingprobleem.* In Hoofdstuk 3, *'Reducing the Optimal to a Useful Number of Clusters for Model Based Clustering'*, worden zes criteria van een goede marktsegmentatie, een informatiecriterium en twee vermoedens die de geometrie van modelgebaseerde clustering modellen beschrijven, gebruikt om het statistisch optimale aantal clusters te reduceren naar een kleiner aantal clusters, geschikt voor de beoogde marketingdoelstellingen.

Zoals eerder genoemd wordt in een groot deel van de literatuur de meest gebruikte clustertechnieken gecontrasteerd. Echter, artikelen die verschillende modelgebaseerde clustertechnieken met elkaar vergelijken (Meila and Heckerman, 2001; Ter Braak et al., 2003) zijn zeldzaam. In Hoofdstuk 4, *'A Comparison of Model Based Clustering Algorithms'*, wordt een vergelijking gemaakt tussen de Bayesiaanse clusterbenadering, zoals beschreven in Hoofdstuk 2, en de benaderingen die zijn geïmplementeerd in LatentGold en Glimmix. Gebruikmakend van simulatiestudies worden de presentaties van deze benaderingen geëvalueerd.

In de eerste drie hoofdstukken ligt de nadruk op de technieken om marktsegmenten of clusters te identificeren. Wanneer deze marktsegmenten of clusters zijn geïdentificeerd, evalueren marketeers de aantrekkelijkheid van elk cluster. De volgende stap in marketing is marktbenadering, wat beschreven is in de laatste twee hoofdstukken. Marktbenadering is het focussen van de marketinginspanningen op de meest winstgevende clusters. Een voorbeeld van zo'n focus is gedifferentieerde marketing. Of, in andere woorden, marketeers proberen exact hetzelfde product of service te verkopen, maar veranderen bijvoorbeeld de promotiemethode voor elk cluster.

Gedifferentieerde marketing als marktbenaderingstrategie kan ook worden geconceptualiseerd met behulp van de heteluchtballon en het voorbeeld in de automarkt. Tijdens de heteluchtballon heb je geleerd dat personen op straat op het ene moment gelijk zijn en op het andere moment verschillend. Je hebt bijvoorbeeld geleerd dat veel personen auto rijden, maar dat de houding ten aanzien van auto's verschillend kunnen zijn tussen groepen van personen. Stel jezelf nu eens voor als een marketeer die auto's moet verkopen. Gebruikmakend van de kennis uit de ballonvaart, kan je verschillende promotiecampagnes ontwikkelen om hetzelfde product, in dit geval auto's, te verkopen. Elke promotiecampagne heeft een verschillende uiting. Een uiting die past bij het te benaderen cluster. Bijvoorbeeld, in een promotiecampagne voor een cluster dat hoofdzakelijk bestaat uit families met jonge kinderen, zal de nadruk van de promotietekst liggen op de veiligheid van de auto. Of op de ruimte van de auto. Voor een cluster dat hoofdzakelijk bestaat uit personen die luxe zoeken, zal de nadruk van de tekst liggen op de luxe componenten van auto's. Of op het aantal paardenkrachten. Verschillende houdingen ten aanzien van auto's moeten leiden tot verschillende promotiecampagnes.

Om clusters zo individueel mogelijk te benaderen, is het belangrijk om zoveel mogelijk van deze personen in deze clusters te leren. Het lijkt dat het verzamelen van de gewenste klantinformatie middels een enkele vragenlijst de beste oplossing is. Echter, doordat tijd en geld gelimiteerd is in de meeste marketingbedrijven, wordt dit vaak niet gerealiseerd. Een aantrekkelijke en praktische oplossing is datafusie, of, in andere woorden, het integreren van verschillende databestanden.

In Hoofdstuk 5, *'The Proof of the Pudding is in the Eating. Data Fusion: An Application in Marketing'*, gepubliceerd als Van Hattum and Hoijtink (2008) en Hoofdstuk 6, *'Improving your Sales with Data Fusion'*, gepubliceerd als Van Hattum and Hoijtink (2009a) wordt getoond hoe de resultaten van twee marktsegmentatiestudies worden gefuseerd met twee klantdatabases. Om het beste datafusie algoritme te selecteren, worden twee traditionele datafusiemethoden, dat zijn polytome logistische regressie en naaste-buren technieken, vergeleken met twee modelgebaseerde clusterbenaderingen. Gebruikmakend van de gefuseerde databestanden worden clusterspecifieke vragenlijsten en clusterspecifieke catalogi gemaakt en verstuurd naar de klanten. De effectiviteit en winstgevendheid van elk datafusie-algoritme wordt vervolgens bepaald door interne en externe criteria.

Het doel van het onderzoek in dit proefschrift was het plaatsen van wetenschappelijk onderzoek over marktsegmentatie en marktbenadering in een bedrijfsperspectief. Om zo realistisch mogelijk te zijn, komen de meeste databestanden uit de dagelijkse marketingpraktijk. Al het onderzoek in dit proefschrift is getest en wordt gebruikt in de dagelijkse bedrijfsvoering van The SmartAgent Company.

Curriculum Vitae

Pascal van Hattum was born on October 22, 1976 in Amersfoort, The Netherlands. He completed pre-university education (VWO) at Farel College in Amersfoort in 1996. He studied Business Mathematics and Computer Science at the VU university Amsterdam from 1996 till 2001. During this study he did an internship at The SmartAgent Company and after graduation he started working at this company as a statistical consultant. In cooperation with The SmartAgent Company he started working part-time on his Ph.D project at the Department Methodology and Statistics at the University of Utrecht in 2004. Currently he is manager Data Intelligence at The SmartAgent Company and combines this function with further research at the Department of Methodology and Statistics at the University of Utrecht.