Joint Modelling of Market Segments and Scale Usage Heterogeneity

Master-Thesis

AUTHOR:

Dawid Piątek

SUPERVISOR:

Prof. Dr. Friedrich Leisch

January 2011

Institut für Statistik,
Ludwig-Maximilians-Universität München

Contents

Chapter 1

Introduction

One kind of data statisticians often face in their work are survey data, where respondents assess their attitude to a subject or express their level of agreement with a statement. Such data are frequently collected using questionnaires containing a set of discrete ratings scales, where respondents have to choose one of a few categories in each question. These scales are often called Likert scales and are commonplace for example in psychology, sociology or market research. From a statistical point of view, this kind of data present two kinds of difficulties. The first is their categorical character and, inherently connected with it, a reduction of information in comparison to standard metric data. The second is a certain level of subjectivity resulting from the self-assessing character of the answers. Despite the fact that all respondents face the same ratings scales, the final interpretation of the ratings, even if categories are labelled, is always left to the person who answers questions. This leads to a situation where the same categories may have different meanings for various persons, which consequently questions the validity of interpersonal comparisons of their answers. Because of possible different interpretations of the meaning of the scales' categories, the use of such ratings scales is much varied - a phenomenon termed a scale usage heterogeneity. The statistical analysis of data affected by scale usage heterogeneity is the subject of my thesis.

The way people use discrete ratings scales with possible biases resulting from it has been an important subject of psychological research for more than sixty years (see e.g. Cronbach, 1946). Paulhus (1991) gives a detailed reference of psychological literature dedicated to this subject, as well as discusses three most prominent response biases: socially desirable responding (SDR), acquiescence response style (ARS), and extreme response style (ERS). Some researches (e.g. Watkins and Cheung, 1995) distinguish between response styles and response sets. The former term refers to a systematical distortion of answers in the way that is independent of the content of questions, the latter to the contamination caused by people's desire to give a particular picture of themselves. Using this distinction, SDR is a response set, whereas ARS and ERS are response styles. In my thesis, I restrict myself to response styles only as, contrary to response sets, they may be accounted for without explicitly taking into account the content of questions. Thus, the terms "response bias" and "response style" will be used interchangeably in the thesis. A more detailed catalogue of possible response styles is presented by Baumgartner and Steenkamp (2001) who mention no less

than seven response styles along with their definitions, theoretical explanations and ways of measuring.

Contrary to the psychological aspects of scale usage heterogeneity, relatively little attention has been devoted to developing statistical methods capable of accounting for differences in response styles. At present, one may differentiate between three not necessarily disjoint, approaches.

The first one assumes that a categorical answer of a respondent results from a discretisation of some latent continuous "true" attitude by a set of thresholds corresponding to categories of the ratings scale. In this case, the response bias affects the way the continuous latent attitude is "translated" into a categorical answer. There are two ways of modeling scale usage heterogeneity using this approach. One is to estimate individual location and scale parameters and, after correcting for them, to use the categorical answers as if they had metrical properties. This allows correcting for most prominent response styles, which are ARS and ERS. Examples of this methodology are (Lenk et al., 2006; Wolfe and Firth, 2002). Alternatively, instead of manipulating distribution of the latent variable, scale usage heterogeneity may be accounted for by allowing heterogeneous thresholds. Depending on the parametrisation of thresholds, such method may be more flexible than the previous one. This methodology was adopted by Johnson (2003). Some authors also combine both methodologies (Rossi et al., 2001; Javaras and Ripley, 2007).

The second approach is based on the item response theory. In this framework, the model predicts the probability of choosing a given category depending on the question being answered and the optional set of regressors. The inclusion of additional variables representing the response style in this set allows one to account for the scale usage heterogeneity. This approach was applied by De Jong et al. (2008), Johnson and Bolt (2010), and Van Rosmalen et al. (2010).

An interesting third approach has been recently proposed by Johnson and Bolt (2010), who use a factor analysis model to identify response styles as additional factors.

In my thesis I analyse scale usage heterogeneity in the context of market segmentation. I examine a method which allows simultaneous classification of respondents with respect to market segments as well as response styles. The method considered in this document follows the first of the aforementioned approaches. I model differences in response styles through heterogeneous thresholds. The reason for that is twofold. Firstly, the use of thresholds offers much greater flexibility in modelling scale usage behaviour than the use of scale and location shift without substantially increasing the number of parameters. Estimation of thresholds gives one the means to model all coherent response styles, i.e. such that preserve the original ordering of categories. For limitations of the scale and location shift methodology in this respect see (Rossi et al., 2001, p. 23). Secondly, using heterogeneous thresholds allows factorisation of scale usage behaviour out of the "true" respondents' attitude in an elegant manner. In this case the whole information about the respondents' attitudes is represented solely by the distribution of the latent variable, whereas the whole scale usage behaviour is represented solely by the thresholds.

This thesis is divided into five chapters. Chapter 1 is introductory, Chapter 2 presents the theory of the considered method. In Chapter 3 I present a set of functions that I wrote in R (R Development Core Team, 2009) for estimating models introduced in Chapter 2, as well as some sample ways of their

use. Results of a simulation study aiming to investigate the performance of the presented method are given in Chapter 4 and a real data example is shown in Chapter 5. Final remarks and a discussion are to be found in Chapter 6.

Chapter 2

Theory

In this chapter I propose a method of a simultaneous estimation of two sorts of clusters: with respect to market segments (hereinafter "opinion clusters" or in short "O-clusters") and scale usage (also referred to as "answer clusters" or in short "A-clusters").

2.1 Statistical model

The presented method assumes that the analysed dataset consists of J categorical variables. Each variable may take one of K values, where 1 represents the least and K the most favourable category (or the weakest and the strongest agreement respectively).

Equations (2.1) to (2.7) present mathematical formulation of the model. Variable x_{ij} represents a categorical response of person i to question j. This takes value k if and only if the corresponding underlying continuous variable z_{ij} takes a value in the interval $(\tau_{a,k-1}, \tau_{a,k}]$. $\tau_{a,k-1}$ and $\tau_{a,k}$ are the two thresholds defining the interval corresponding to category k. Since there are K possible categories, there must be $K + 1$ thresholds. The thresholds' values depend on A-cluster the person belongs to, but are common across all questions. Vector $\mathbf{z}_i$ represents the respondent's i attitude to all questions and is assumed to follow multivariate normal distribution with parameters defined by the O-cluster he or she belongs to. The parentheses around the subscript by the covariance matrix in (2.2) indicate that it may or may not be O-cluster specific. This is analogous to the situation in finite mixture modelling (see e.g. Celeux and Govaert, 1995; Fraley and Raftery, 2002, for more details). Equations (2.3) to (2.5) state that there are K categories, A answer clusters and O opinion clusters. Because $\mathbf{z}$ is assumed to follow the normal distribution, both most extreme thresholds are fixed at infinities.

$$x_{ij} = k \quad \Leftrightarrow \quad \tau_{a,k-1} < z_{ij} \leq \tau_{a,k} \tag{2.1}$$
$$\mathbf{z}_i \sim N(\mu_o, \Sigma_{(o)}) \tag{2.2}$$

$$k = 1 \ldots K \tag{2.3}$$
$$a = 1 \ldots A \tag{2.4}$$
$$o = 1 \ldots O \tag{2.5}$$
$$\tau_{a,0} = -\infty \tag{2.6}$$
$$\tau_{a,K} = \infty \tag{2.7}$$

2.2 Estimation

The most fundamental idea behind the model formulated above is that heterogeneity of the whole analysed population may be accurately summarised with respect to both the represented opinions and exhibited scale usage behaviour by a moderate number of homogeneous clusters. If this is true, we may be interested in identifying these clusters and utilising the information gained in one sort of clustering in order to improve the quality of the other sort of clustering and vice versa. To achieve such mutual reinforcement of both sorts of clusters, I propose an iterative procedure described in Algorithm 1.

- Choose initial partition into O-clusters and latent distribution for every O-cluster.
- Choose initial partition into A-clusters and thresholds' values in every A-cluster.
- Repeat until a satisfying solution is found or the maximum number of iterations is reached:
  1. Find O-clusters conditional to the current thresholds' estimates.
  2. Estimate distribution parameters of the O-clusters found.
  3. Find A-clusters conditional to the current distribution estimates.
  4. Estimate thresholds for the A-clusters found.

Algorithm 1: Estimation algorithm for the presented model.

Algorithm 1 is very general so in order to be applicable it needs further specification. First of all, clustering procedures for both types of clustering must be chosen. Some clustering procedures may require computing expected values of latent variables $\mathbf{z}$. This is done within step 1. The two steps before the main loop of the algorithm allow the set up of initial partitions, if required by chosen clustering procedures. Any existing prior knowledge regarding one or both partitions can be used here. If no prior knowledge is available, all the cases will typically be classified into a single cluster. Algorithm 1 is not guaranteed to converge, hence a very general exit condition for the main loop of the algorithm.

In the following I present specifications of Algorithm 1, which I examined in detail. They differ with respect to the clustering procedures applied and the way they handle the categorical nature of data. A natural way of dealing with categorical character of data seems to be a latent version of the model-based clustering presented in Section 2.2.1. Due to a large computational burden connected with this procedure, one may be interested in simpler and faster alternatives presented in Section 2.2.2.

2.2.1 Latent model-based O- & hierarchical A- clustering

The approach presented in this section proposes a latent version of model-based clustering for O-clustering (see e.g. Fraley and Raftery, 2002, for a review). It gives the advantage of direct accounting for the categorical character of data. Large computational burden connected with computing O-clusters and the fact that, as a consequence, the number of O-clusters must be chosen practically in advance are its shortcomings.

Initialisation Latent model-based O-clustering does not require initial O-partition, so only preliminary A-clusters need to be specified. If no prior knowledge regarding the scale usage is available in a given sample, a single A-cluster with thresholds equal to the quantiles of the standard normal distribution is a natural choice. This fixes the location and scale of estimated O-clusters' latent distributions and allows to relate them to the standard normal. In this case, 0 value of latent variable z corresponds to the middle of the middle category (if the number of categories is odd) or to the value of the threshold between the two middle categories (if the number of categories is even). Variance is fixed in such a way that when the expected value is 0, a unit variance in a given O-cluster means equal probability of selecting any of K categories. Once the initial A-partition is set, the main loop of Algorithm 1 may be entered.

Steps 1 and 2 In model-based clustering, the optimal partition and clusters' parameters are estimated simultaneously, so that steps 1 and 2 of Algorithm 1 are merged. The analysed population is assumed to follow a finite mixture of multivariate normal distributions, every component of which corresponds to one O-cluster. Model-based clusters are estimated using the EM algorithm (Dempster et al., 1977), which consists of two steps: an expectation and a maximisation step. In the former a (fuzzy) partition into clusters is estimated. In the latter clusters' parameters as well as mixture proportions are determined.

In order to apply the EM algorithm, the likelihood function needs to be specified. We start with defining a conditional probability that person i, who belongs to A-cluster a and to O-cluster o, in response to J questions gives a categorical vector $\mathbf{k}$:

$$\pi_{i|ao} \triangleq P(\mathbf{x}_i = \mathbf{k}|\boldsymbol{\mu}_o, \Sigma_o, \boldsymbol{\tau}_a) = \int_{\tau_{a,k(1)-1}}^{\tau_{a,k(1)}} \int_{\tau_{a,k(2)-1}}^{\tau_{a,k(2)}} \dots \int_{\tau_{a,k(J)-1}}^{\tau_{a,k(J)}} dN(\mu_o, \Sigma_o). \tag{2.8}$$

Notation $\int dN(\mu_o, \Sigma_o)$ represents the integral over the density function of the normal distribution with parameters μ_o and Σ_o, whereas $k(j)$ represents the j-th element of vector $\mathbf{k}$. Vectors $\mathbf{k}$ will be also referred to as "response patterns"

and indexed with r. Therefore, the above probability will also sometimes be denoted $\pi_{r|ao}$ when it refers to a particular response pattern r instead of a person i.

A-clusters are defined and fixed during O-clustering. Fuzzy A-clusters are not allowed, so every person belongs to only one A-cluster. It is assumed that A- and O-clusters are independent, so that A-cluster membership does not directly influence O-cluster membership. Furthermore, every response pattern r has an unambiguously specified A-cluster it belongs to. Because of this, the only way how the A-clusters must be accounted for during O-clustering is by using appropriate thresholds $\boldsymbol{\tau}_a$ in computing $\pi_{i|ao}$.

The probability that a person i belongs to the o-th component of the mixture density and responds a vector $\mathbf{k}$ given that they belong to the A-cluster a, is:

$$P(\mathbf{x}_i = \mathbf{k}|\lambda_o, \boldsymbol{\mu}_o, \Sigma_o, \boldsymbol{\tau}_a) = \lambda_o \pi_{i|ao}, \tag{2.9}$$

where λ_o is a mixture proportion of component o in the mixture density.

Let us introduce an additional set of O variables γ_{io} which represent the O-cluster membership of the person i. If the person i belongs to the O-cluster o, the o-th of these variables equals 1 and the rest are 0s, hence (2.9) may be rewritten as:

$$P(\mathbf{x}_i = \mathbf{k}|\lambda_o, \boldsymbol{\mu}_o, \Sigma_o, \boldsymbol{\tau}_a) = \prod_o (\lambda_o \pi_{i|ao})^{\gamma_{io}}. \tag{2.10}$$

In the above equation $\prod_o$ denotes the product over all O-clusters. An analogous notation for products and sums will be used throughout the rest of the thesis. Equation (2.10) leads to the following (log-)likelihood function[1]:

$$L = \prod_i \prod_o (\lambda_o \pi_{i|ao})^{\gamma_{io}}, \tag{2.11}$$

$$l = \ln L = \sum_i \sum_o \gamma_{io} \ln(\lambda_o \pi_{i|ao}), \tag{2.12}$$

where

$$\sum_o \gamma_{io} = 1 \quad \wedge \quad \forall \gamma_{io} \geq 0. \tag{2.13}$$

Having the above log-likelihood function in mind, we may specify the two steps of the EM algorithm (for details see e.g. McLachlan and Peel, 2000, sec. 2.8).

In the E step the γ_{io}'s are estimated using their conditional expected values:

$$\gamma_{io} = \frac{\lambda_o \pi_{i|ao}}{\sum_{u \in O} \lambda_u \pi_{i|au}}. \tag{2.14}$$

In the M step λ_o's are estimated as mean values of γ_{io}'s in a given O-cluster:

$$\lambda_o = \frac{1}{N} \sum_i \gamma_{io}, \tag{2.15}$$

[1] In the literature dedicated to the EM algorithm this likelihood is also called "complete data likelihood" (for details see e.g. McLachlan and Peel, 2000, sec. 1.9).

and the parameters of O-clusters μ_o and Σ_o are estimated by likelihood maximisation. However, a direct maximisation of the log-likelihood given in equation (2.12) requires integration over multivariate normal density function, dimensionality of which is equal to the number of questions. This is computationally feasible only when the number of variables is small (Lee et al., 1990; Jöreskog and Moustaki, 2001). In order to estimate parameters of the mixture density's components, I use the "Underlying Bivariate Normal" (UBN) approach proposed by Jöreskog and Moustaki (2001). This is a limited information maximum likelihood method. Instead of maximising the full likelihood function of J dimensional distribution, it maximises the sum of univariate and bivariate marginal likelihoods only. The univariate marginal probability that a person i from A-cluster a and O-cluster o responds k to question j is:

$$\pi^{(j)}_{k|ao} \triangleq P(x_{ij} = k|\mu_{oj}, \sigma^2_{oj}, \boldsymbol{\tau}_a) = \int_{\tau_{a,k-1}}^{\tau_{a,k}} dN(\mu_{oj}, \sigma^2_{oj}). \tag{2.16}$$

The analogous bivariate probability that a person i responds k to question j and m to question l is:

$$\begin{aligned} \pi^{(jl)}_{k,m|ao} &\triangleq P(x_{ij} = k \wedge x_{il} = m \mid \mu_{oj}, \mu_{ol}, \sigma^2_{oj}, \sigma^2_{ol}, \rho_{jl}, \boldsymbol{\tau}_a) \\ &= \int_{\tau_{a,k-1}}^{\tau_{a,k}} \int_{\tau_{a,m-1}}^{\tau_{a,m}} dN(\boldsymbol{\mu}_{jl}, \Sigma_{jl}), \end{aligned} \tag{2.17}$$

where

$$\boldsymbol{\mu}_{jl} \triangleq \begin{bmatrix} \mu_{oj} \\ \mu_{ol} \end{bmatrix}, \quad \Sigma_{jl} \triangleq \begin{bmatrix} \sigma^2_{oj} & \rho_{jl}\sigma_{oj}\sigma_{ol} \\ \rho_{jl}\sigma_{oj}\sigma_{ol} & \sigma^2_{ol} \end{bmatrix}. \tag{2.18}$$

The sum of all univariate and bivariate log-likelihoods for a model-based O-cluster o has the following form:

$$\begin{aligned} l^{(\mathrm{UBN})}_o = \sum_a \dot{p}_{a|o} \Bigg(& \sum_{j=1}^{J} \sum_{k=1}^{K} \dot{p}^{(j)}_{k|ao} \ln \pi^{(j)}_{k|ao} + \\ & \sum_{j=2}^{J} \sum_{l=1}^{j} \sum_{k=1}^{K} \sum_{m=1}^{K} \dot{p}^{(jl)}_{k,m|ao} \ln \pi^{(jl)}_{k,m|ao} \Bigg), \end{aligned} \tag{2.19}$$

where

$$\dot{p}_{a|o} \triangleq \frac{\sum_{i \in I(a)} \gamma_{io}}{\sum_i \gamma_{io}}, \tag{2.20}$$

$$\dot{p}^{(j)}_{k|ao} \triangleq \frac{\sum_{i \in \{I(a) \cap I(x_{ij}=k)\}} \gamma_{io}}{\sum_{i \in I(a)} \gamma_{io}}, \tag{2.21}$$

$$\dot{p}^{(jl)}_{k,m|ao} \triangleq \frac{\sum_{i \in \{I(a) \cap I(x_{ij}=k) \cap I(x_{il}=m)\}} \gamma_{io}}{\sum_{i \in I(a)} \gamma_{io}}. \tag{2.22}$$

In the above equations, $I()$ denotes a set of respondents for whom the condition in brackets is true, whereas $I(a)$ denotes a set of respondents belonging to the

A-cluster a. The $\dot{p}$ terms may be interpreted respectively as a proportion of weights γ_{io} of cases belonging to the A-cluster a among the members of the O-cluster o, proportion of weights γ_{io} of cases who answered k to question j in the interception of the A-cluster a and the O-cluster o and proportion of weights γ_{io} of cases who answered k to question j and m to question l in the interception of the A-cluster a and the O-cluster o.

Iterating E and M steps until convergence leads to a partition of the dataset into O-clusters and delivers estimates of parameters of every O-cluster. The convergence of the EM algorithm was proved in the seminal paper by Dempster et al. (1977).

If the model were estimated using the usual log-likelihood function, the quality of the fit could be evaluated using the value of the log-likelihood function given by (2.12). However, since I use the function given by (2.19) instead, I derive an analogous criterion using this function. Hence, equation (2.12) may be transformed to the following form:

$$l = \sum_i \sum_o \gamma_{io} \ln \lambda_o + \sum_i \sum_o \gamma_{io} \ln \pi_{i|ao}. \tag{2.23}$$

The two above terms have clear interpretations. The former is responsible for estimating the mixture proportions (its differentiation with respect to λ_o's leads to (2.15)), the latter is responsible for the parameters of particular O-clusters. In the UBN approach, instead of the latter term I maximise the expression given by (2.19). Thus, a natural approach is to use analogous substitution in the fit function. As a consequence, the UBN equivalent of the log-likelihood, which can be used to assess the model fit, has the following form:

$$\begin{aligned} l^{(\mathrm{UBN})} = & \sum_o \ln \lambda_o \sum_i \gamma_{io} + \\ & \sum_o \sum_a \dot{p}_{a|o} \left(\sum_{j=1}^{J} \sum_{k=1}^{K} \dot{p}^{(j)}_{k|ao} \ln \pi^{(j)}_{k|ao} + \sum_{j=2}^{J} \sum_{l=1}^{j} \sum_{k=1}^{K} \sum_{m=1}^{K} \dot{p}^{(jl)}_{k,m|ao} \ln \pi^{(jl)}_{k,m|ao} \right). \end{aligned} \tag{2.24}$$

Step 3 Once the O-partition and the O-clusters are estimated, the data are clustered according to scale usage pattern. In my thesis I examine three approaches for finding optimal A-partitions. Here, the all three are combined with hierarchical clustering, but in principle, any standard clustering method could be used instead. The first step for each of these approaches is computing individual thresholds $\boldsymbol{\tau}_i$. They are estimated by maximising individual likelihoods given by (2.10) with respect to a vector of individual thresholds $\boldsymbol{\tau}_i$, which replaces A-cluster specific thresholds $\boldsymbol{\tau}_a$. In order to reduce the computational burden, a simplifying assumption is made that the covariance matrices in all O-clusters are diagonal. This results in ignoring all the correlations computed in step 2. Thus, the conditional probability of a response pattern r $\pi_{r|ao}$ may be computed simply as a product of univariate marginal probabilities:

$$\pi_{r|ao} = \prod_j \pi^{(j)}_{k(j)|ao}, \tag{2.25}$$

where $\pi^{(j)}_{k(j)|ao}$ is defined as in (2.16).

Logarithming (2.10) utilising (2.25) leads to the following log-likelihood:

$$\ln\Big(P(\mathbf{x}_i = \mathbf{k}|\lambda_o, \boldsymbol{\mu}_o, \Sigma_o, \boldsymbol{\tau}_i)\Big) \tag{2.26}$$

$$= \ln\Big(\prod_o (\lambda_o \pi_{i|ao})^{\gamma_{io}}\Big) \tag{2.27}$$

$$= \sum_o \gamma_{io} \ln\Big(\lambda_o \prod_j \pi^{(j)}_{k(j)|ao}\Big) \tag{2.28}$$

$$= \underbrace{\sum_o \gamma_{io} \ln \lambda_o}_{\text{const}(\boldsymbol{\tau}_i)} + \sum_o \gamma_{io} \sum_j \ln \pi^{(j)}_{k(j)|ao} \tag{2.29}$$

$$\propto \sum_o \gamma_{io} \sum_j \ln \pi^{(j)}_{k(j)|ao}. \tag{2.30}$$

Having the individual thresholds computed, the three aforementioned approaches differ in the way these thresholds are utilised. In the first approach, the thresholds are directly clustered, which has the following drawback: in cases where respondents did not choose some of the extreme categories, the estimates of the corresponding thresholds tend to have large absolute values. This may result in the respondents being divided into groups according to irrelevant criteria. To avoid such influence, before applying clustering, I truncate all thresholds' values at 4. In the second approach, I apply estimated thresholds to the standard normal distribution to compute probabilities for each category and use these probabilities in clustering. In the third approach, along with the previously mentioned probabilities, a measure of acquiescence is computed and added to probabilities while clustering. The measure of acquiescence is computed in the following way: first, the categories are renumbered so that the middle category (or the neutral, if they do not coincide) has 0 value, positive categories have the following positive integers, and the negative categories analogous negative values. Then, the values for all categories are multiplied by corresponding probabilities and summed. This may be interpreted as the expected value of the categorical answer after renumbering under the probability distribution induced by the estimated thresholds and the underlying standard normal distribution. The rationale for all the three approaches is given in Section 4.3.1 ,which presents the results of clustering using expected values of latent variables on simulated data.

Independently of the approach used, the result of this step is a new partition of data with respect to the scale usage behaviour, which is then used in step 4.

Step 4 In step 4 new thresholds for every newly found A-cluster are estimated by maximising the same log-likelihood as in (2.30), but over all members of the A-cluster and with respect to A-cluster thresholds $\boldsymbol{\tau}_a$. The (log-)likelihood function for the A-cluster a has the following form:

$$L_a = \prod_{i \in I(a)} \prod_o (\lambda_o \pi_{i|ao})^{\gamma_{io}}, \tag{2.31}$$

$$l_a = \sum_{i \in I(a)} \sum_o \gamma_{io} \ln \lambda_o + \sum_{i \in I(a)} \sum_o \gamma_{io} \sum_j \ln \pi^{(j)}_{k(j)|ao} \tag{2.32}$$

$$\propto \sum_{i \in I(a)} \sum_o \gamma_{io} \sum_j \ln \pi^{(j)}_{k(j)|ao}. \tag{2.33}$$

The full form of the log-likelihood function in (2.32) may be used to assess the fit of A-clusters.

Once the new estimates of distributions within O-clusters and new threshold estimates for every A-cluster are computed, the next iteration of the main loop may begin.

2.2.2 Clustering using expected values of latent variables

This approach is a way to overcome computational problems connected with the latent model-based clustering presented in Section 2.2.1. Instead of using computationally demanding probabilities of hyperrectangular cut-outs of the multivariate normal distribution, an expected value of latent variables $\mathbf{z}$ for every hyperrectangle is computed and used in classical clustering procedures. For the sake of computational simplicity, independence of all variables is assumed analogous to step 3 of the previous section.

Theoretically, any clustering procedure can be used. Here, I consider hierarchical clustering when the number of clusters is unknown, and k-means clustering when the number of clusters is known.

Initialisation At the beginning, all cases are classified into a single O-cluster for which multivariate spherical standard normal distribution is assumed. Next, all cases are classified into a single A-cluster and thresholds corresponding to the quantiles of the standard normal distribution are assumed.

Step 1 In step 1 expected values of the latent variables $\mathbf{z}$ are computed for every person conditional on the A- and O- cluster the person belongs to. Due to the independence assumption, this may be done separately for every univariate z_j. Then, these expected values are clustered in a standard way using e.g. hierarchical or k-means clustering.

Step 2 In step 2 parameters of the latent distribution are computed for every O-cluster using the maximum likelihood method. The (log-)likelihood for the O-cluster o is:

$$L_o = \prod_a \prod_r \pi_{r|ao}^{n_{aor}}, \tag{2.34}$$

$$l_o = \ln L_o = \sum_a \sum_r n_{aor} \ln \pi_{r|ao} \tag{2.35}$$

$$= n_o \sum_a \sum_r p_{ar|o} \ln \pi_{r|ao} \tag{2.36}$$

$$\propto \sum_a p_{a|o} \sum_r p_{r|ao} \ln \pi_{r|ao}. \tag{2.37}$$

In the above equations n_{aor} denotes the number of answered response patterns r in the intersection of the A-cluster a and the O-cluster o, n_o is the number of persons in the O-cluster o, $p_{ar|o}$ is the proportion of response patterns r given in the A-cluster a in relation to the size of the O-cluster o (i.e. n_{aor}/n_o), $p_{a|o}$ is the proportion of members of the A-cluster a in the O-cluster o (n_{ao}/n_o) and $p_{r|ao}$ is the proportion of the number of response patterns r in the intersection of appropriate A- and O- clusters.

Step 3 In step 3 individuals are A-clustered in similar manner as in Section 2.2.1, i.e. the optimal individual thresholds are estimated first and then, depending on the chosen approach, appropriate values are clustered using standard procedures, such as hierarchical or k-means clustering. Since, contrary to the latent model-based O-clusters, we have hard O-clusters and the latent variables z_j are assumed to be independent, the individual log-likelihood is simply the product of j marginal probabilities from equation (2.16) with $\boldsymbol{\tau}_a$ replaced by $\boldsymbol{\tau}_i$.

Step 4 In step 4 A-cluster specific thresholds are estimated by maximising likelihood in an analogous manner as used in step 2, but with respect to the vector of thresholds $\boldsymbol{\tau}_a$:

$$L_a = \prod_o \prod_r \pi_{r|ao}^{n_{aor}}, \tag{2.38}$$

$$l_a = \ln L_a = \sum_o \sum_r n_{aor} \ln \pi_{r|ao} \tag{2.39}$$

$$= n_a \sum_o \sum_r p_{ar|o} \ln \pi_{r|ao} \tag{2.40}$$

$$\propto \sum_o p_{o|a} \sum_r p_{r|ao} \ln \pi_{r|ao}. \tag{2.41}$$

Once new thresholds are estimated, a new iteration of the main loop may begin.

Chapter 3

Implementation

This chapter describes implementation of the methods presented in Chapter 2. Due to general character of Algorithm 1, these methods were implemented as a set of functions in R language (R Development Core Team, 2009). This allows flexible construction of the main loop of Algorithm 1 using various functions depending on the chosen building blocks of the algorithm. In this chapter I present and describe examples of the main loop of the algorithm for both variants presented in Sections 2.2.1 and 2.2.2 constructed with a set of functions I implemented. In all listings ,these functions are coloured violet; for the sake of simplicity, in all listings `X` represents the data matrix containing categorical responses on the scale from 1 to 5.

3.1 Latent model-based O- & hierarchical A- clustering

Listing 3.1 presents a sample code for the procedure described in Section 2.2.1. In this variant, O-clustering is performed using latent model-based clustering, whereas A-clustering is carried out using hierarchical clustering of thresholds.

```
###
### Latent model-based O- and hierarchical A- clustering.
###

iters <- list()

a.part <- rep(1, nrow(X))
aclust <- list(thrs=matrix(qnorm(1:4/5), nrow=1))
no <- 4
oclust <- list(p.mat=v2mx(rep(1:no,length=nrow(X)),no))

for (i in 1:10) {
    print(paste("Iteration",i))

    ## Steps 1,2
    ## Latent model-based O-clustering:
    oclust <- latent.Mclust(data=X, no=no, aclust=a.part,
        aclust.thrs=aclust$thrs, p.mat=oclust$p.mat)
    fit.value(lMclust=oclust)

    ## Step 3
```

```
21      ## Estimating individual thresholds given O-clusters
        ind.thrs <- estimate.ind.thresholds(data=X,
            oclust.pars=oclust$o.pars, p=oclust$p.mat)
23      pairs(ind.thrs)

25      ## A-clustering
        ind.thrs.r <- apply(ind.thrs,c(1,2), to.range,4)
27      ad <- dendrogram(ind.thrs.r)
        a.part <- clusters(3,ad)
29      abarplots(data=X, aclust=a.part)

31      ## Step 4
        ## Computing thresholds for A-clusters
33      aclust <- estimate.cluster.thresholds(data=X, aclust=a.part,
            oclust.pars=oclust$o.pars, p=oclust$p.mat)
        fit.value(lMclust=oclust,aclust=aclust)
35      plot.thrs(aclust$thrs, c=2)

37      ## Saving iteration:
        iters[[i]] <- list(oclust=oclust, ind.thrs=ind.thrs, ad=ad,
            a.part=a.part, aclust=aclust)
39  }
```

Listing 3.1: A sample code for latent model-based O- & hierarchical A-clustering.

In the first step in Listing 3.3, before the algorithm begins, a list `iters` is created. It holds all objects created in every iteration, which allows tracing of the course of the algorithm.

Initialisation The algorithm starts with setting up an initial A-partition. Here, all observations are classified into a single A-cluster with thresholds corresponding to the quantiles of the standard normal distribution. Latent model-based O-clustering requires some starting values for the EM algorithm used for estimating O-clusters. This may be the initial partition of data or parameters of the mixture distribution. In this case, I set the number of O-clusters in the variable `no` to 4 and define a starting partition using a matrix of γ_{io}'s. Function `v2mx` in line 10 converts the vector of repeating 1 to 4 sequences into a matrix with the number of rows equal to the length of the vector and four columns. In every row of the matrix a 1 is put in a column indicated by the value of the corresponding vector's entry and the remaining entries of the matrix are filled with 0s. In this way every person is classified into one of the O-clusters and, if the number of cases is a multiple of the number of assumed O-clusters, all clusters have equal size.

Steps 1 and 2 As mentioned in Section 2.2.1 steps 1 and 2 are merged. The whole latent model-based clustering is performed by function `latent.Mclust`. The required input in this function is a data matrix, the number of O-clusters to estimate, parameters of the A-clusters and a starting point for the EM algorithm - in this example the matrix defined in the initialisation step.

The object returned by `latent.Mclust` function may be used to assess the quality of fit by function `fit.value`, which evaluates the fit function in (2.24).

Step 3 In step 3, individual thresholds are estimated by maximising the log-likelihood function given by (2.30) for every person. This is done by function

`estimate.ind.thresholds`. These individual thresholds can be visualised using e.g. `pairs` function. Extreme values of the "outer" thresholds are a frequently seen pattern, which may lead to irrelevant partitions from the scale usage point of view. To avoid this, all thresholds' values are truncated to interval [-4,4] by applying function `to.range` to every entry of matrix `ind.thrs`. The result is a matrix of truncated thresholds `ind.thrs.r`, used for actual A-clustering. This is done using `dendrogram` function, which performs hierarchical clustering and displays the resulting dendrogram in order to facilitate the choice of the number of clusters. When the number of clusters is chosen, partitioning of data is done using function `clusters`, which is merely a wrapper for the standard R function `cutree`. Function `abarplots` displays barplots showing the distribution of the categories in all created A-clusters.

To use a different method of A-clustering rather than direct clustering of thresholds, one must change lines 25 - 29 in Listing 3.1. Listing 3.2 presents a sample code for A-clustering using induced probabilities and the ARS measure. First, individual thresholds are transformed into induced probabilities using function `tau2prob`. Then, the ARS measure is computed. Since there are 5 categories and the rating scale is symmetric, new categories span from -2 to 2. Finally, `dendrogram` function is applied to the probabilities combined with the ARS measure. Since the ARS measure has a different scale than the probabilities, it is scaled by a factor of 0.6. The rationale for using this particular value is given in Section 4.3.1. The rest of the code remains unchanged.

```
    ## A-clustering using induced probabilities
2   ## and the ARS measure.
    ind.thrs.p <- tau2prob(ind.thrs)
4   ars <- apply(ind.thrs.p,1,function(x) sum((-2:2)*x))
    ad <- dendrogram(cbind(ind.thrs.p,0.6*ars))
6   a.part <- clusters(3,ad)
    abarplots(data=X, aclust=a.part)
```

Listing 3.2: A sample code for A-clustering using induced probabilities and the ARS measure.

Step 4 In this step, thresholds $\boldsymbol{\tau}_a$ are estimated for every A-cluster. Function `estimate.cluster.thresholds` maximises thelog-likelihood given in (2.33) using the A-partition from step 3 and thecurrent O-partition. Function `fit.value` in line 34 evaluates the log-likelihood function given by (2.32) to assess the quality of fit. Estimated thresholds may be visualised using function `plot.thrs`.

Finally, all objects created in the current iteration are saved in the `iters` list and a new iteration begins.

3.2 Clustering using expected values of latent variables

Listing 3.3 presents a code for the faster procedure using expected values of latent variables described in Section 2.2.2. To illustrate both the k-means and hierarchical clustering, I use the former for O-clustering and the latter for A-clustering.

Similarly to the previous section, I start with creating a list which will store all important objects created during the course of the algorithm.

Initialisation Contrary to the previous section, both types of partitions must be fully specified in the initial step, so that the expected values could be computed. Here, I classify all observations into single A- and O- clusters. I choose standard normal quantiles as thresholds for the A-cluster. The parameters of the initial O-cluster are estimated using function `distributions.within.oclusters`, described along with function `plot.oclust` in step 2, where they are typically used.

Step 1 Step 1 starts with computing expected values of the latent variables `z`. This is done by function `compute.latent.variables`. Since all latent variables z_j are assumed to be independent, as explained in Section 2.2.2, each univariate expected value is computed separately . Subsequently, all are combined in a vector. These vectors are computed for every person in a sample and form a matrix denoted with `z` in Listing 3.3. The expected values are clustered in a standard way using the `kmeans` function to find four clusters using 5 different starting values. The best partition found is saved as `o.part`. Function `oboxplots` offers a visualisation of newly estimated O-clusters.

Step 2 A multivariate normal distribution is assumed for every O-cluster found in step 1. Parameters of each such distribution are estimated using function `distributions.within.oclusters`, which maximises the log-likelihood in (2.37). Function `plot.oclust` offers visualisation of O-clusters. Quality of fit may be assessed using function `fit.value`, which, given the output from `distributions.within.oclusters`, returns the value of the full log-likelihood function as defined in (2.36).

Step 3 A-clusters are built in a similar manner to the one described in Section 4.2. First, individual thresholds are estimated using function `estimate.ind.thresholds`. The only difference is that instead of (2.30), it maximises (2.41). Then, like in Section 4.2, these thresholds are truncated to the interval [-4,4] and clustered hierarchically. Alternatively, induced probabilities with or without the ARS measure may be used in exactly the same manner as in Listing 3.2.

Step 4 Step 4 goes exactly as in Section 4.2. Function `estimate.cluster.thresholds` maximises for every A-cluster (2.41) to estimate A-cluster specific thresholds. In this case, function `fit.value` evaluates (2.40) and `plot.thrs` offers a visualisation of the thresholds.

After all the relevant objects are saved in `iters`, a new iteration begins.

```
###
### K-means O- and hierarchical A- clustering:
###

iters <- list()

a.part <- rep(1, nrow(X))
o.part <- rep(1, nrow(X))
aclust <- list(thrs=matrix(qnorm(1:4/5), nrow=1))

## Initial global distribution
oclust <- distributions.within.oclusters(data=X, ncat=5,
    oclust=o.part, aclust.thrs=aclust$thrs, aclust=a.part)
plot.oclust(oclust$pars, 1, ncat=5, r=3, c=3)

for (i in 1:10) {
    print(paste("Iteration",i))

    ## Step 1
    ## estimating expected values of latent variables z
    z <- compute.latent.variables(data=X, oclust.pars=oclust$pars,
        oclust=o.part, aclust.thrs=aclust$thrs, aclust=a.part)

    ## O-clustering
    okm <- kmeans(z,centers=4,nstart=5)
    o.part <- okm$cluster
    oboxplots(z, oclust=o.part, c=2)

    ## Step 2
    ## Estimating parameters of the latent distributions
    ## within O-clusters
    oclust <- distributions.within.oclusters(data=X,
        oclust=o.part, aclust.thrs=aclust$thrs, aclust=a.part,
        ncat=5)
    for (o in 1:4) plot.oclust(oclust$pars, o, ncat=5, r=3, c=3)
    fit.value(oclust=oclust)

    ## Step 3
    ## Estimating individual thresholds given O-clusters
    ind.thrs <- estimate.ind.thresholds(data=X,
        oclust.pars=oclust$pars, oclust=o.part)
    pairs(ind.thrs)

    ## A-clustering
    ind.thrs.r <- apply(ind.thrs, c(1,2), to.range, 4)
    ad <- dendrogram(ind.thrs.r)
    a.part <- clusters(3, ad)
    abarplots(data=X, aclust=a.part)

    ## Step 4
    ## Computing thresholds for A-clusters
    aclust <- estimate.cluster.thresholds(data=X, aclust=a.part,
        oclust.pars=oclust$pars, oclust=o.part)
    fit.value(aclust=aclust)
    plot.thrs(aclust$thrs, c=2)

    ## Saving iteration
    iters[[i]] <- list(z=z, od=od, o.part=o.part, oclust=oclust,
        ad=ad, a.part=a.part, aclust=aclust)
}
```

Listing 3.3: Sample code for clustering using expected values of latent variables.

Chapter 4

Simulated data results

In this chapter I present the results of applying the methods described in Chapter 2 to simulated data. In the first section I describe simulated data. In the following sections I present the results of applying the methods described in Sections 2.2.1 and 2.2.2 to these data.

4.1 Data

To test the methods presented in Chapter 2, I simulated a dataset consisting of nine categorical variables, each taking one of five values between 1 and 5. The data were generated in two steps. First, latent variables $\mathbf{z}$ were simulated. 300 observations were drawn from each of 4 multivariate normal distributions with different mean vectors, to get a dataset presented in Figure 4.1. As can be seen in the figure, the nine variables form three blocks. The variables are strongly correlated within the blocks, but there are no correlations between the blocks. Then, every hundred in each O-cluster was categorised using thresholds corresponding to one of three A-clusters. The thresholds used for every of these A-clusters are shown in Table 4.2 and plotted in Figure 4.2. Additionally, Figure 4.3 depicts distributions of answered categories in every A-cluster. The effect of varying thresholds on categorical data is illustrated in Figure 4.4. In this way a dataset of 1200 observations is created, one hundred for every intersection of the A- and O- clusters.

O-cluster	Mean vector μ_o^T								
O1	[0	0	0	0.7	0.7	0.7	0.7	0.7	0.7]
O2	[0.7	0.7	0.7	-0.7	-0.7	-0.7	0	0	0]
O3	[-0.7	-0.7	-0.7	0.7	0.7	0.7	-0.7	-0.7	-0.7]
O4	[-0.7	-0.7	-0.7	0	0	0	0.7	0.7	0.7]

Table 4.1: Vectors of means of four simulated O-clusters.

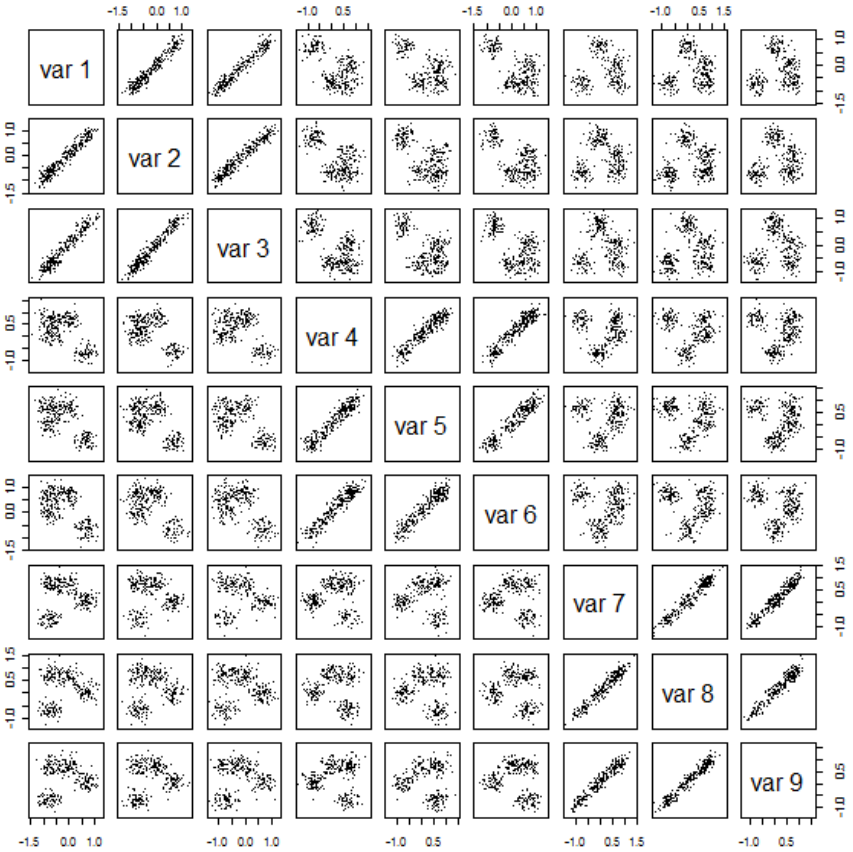


Figure 4.1: Scatterplot matrix of simulated latent variables.

A-cluster	Vector of thresholds $\boldsymbol{\tau}_a^T$							
A1	[	$-\infty$	-0.84	-0.25	0.25	0.84	∞	]	
A2	[	$-\infty$	-1.34	-0.75	-0.25	0.34	∞	]	(ARS)
A3	[	$-\infty$	-0.25	-0.2	0.2	0.25	∞	]	(ERS)

Table 4.2: Thresholds used in three simulated A-clusters.

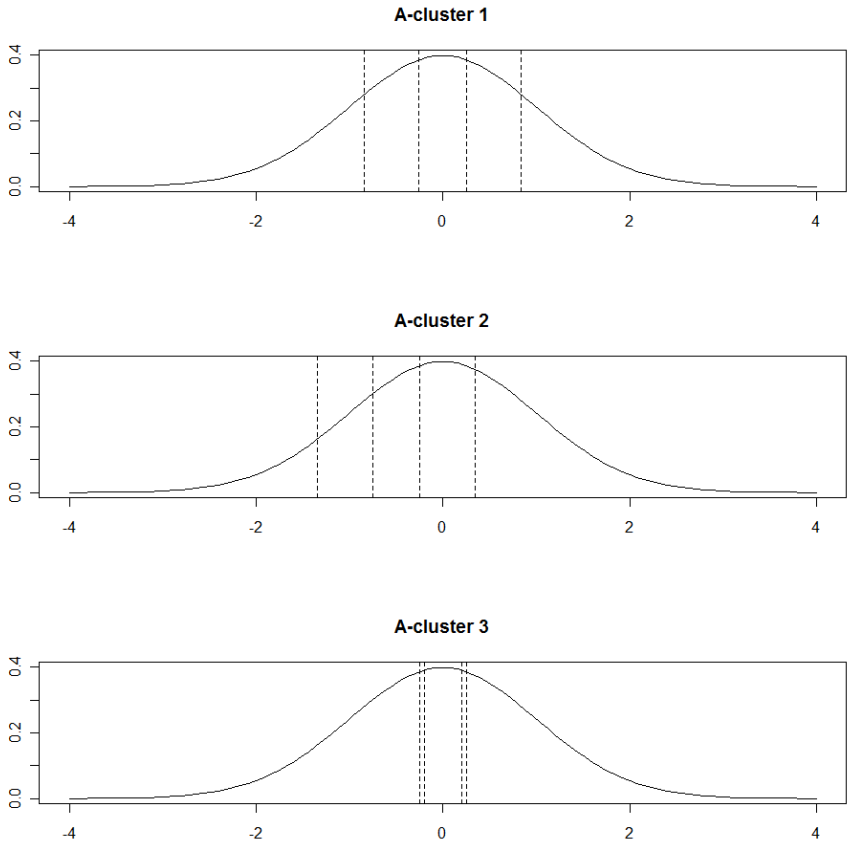


Figure 4.2: Thresholds used in three simulated A-clusters.

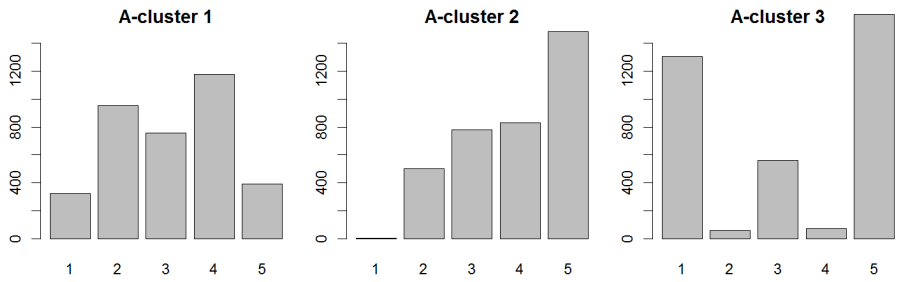


Figure 4.3: Distributions of answers within true A-clusters.

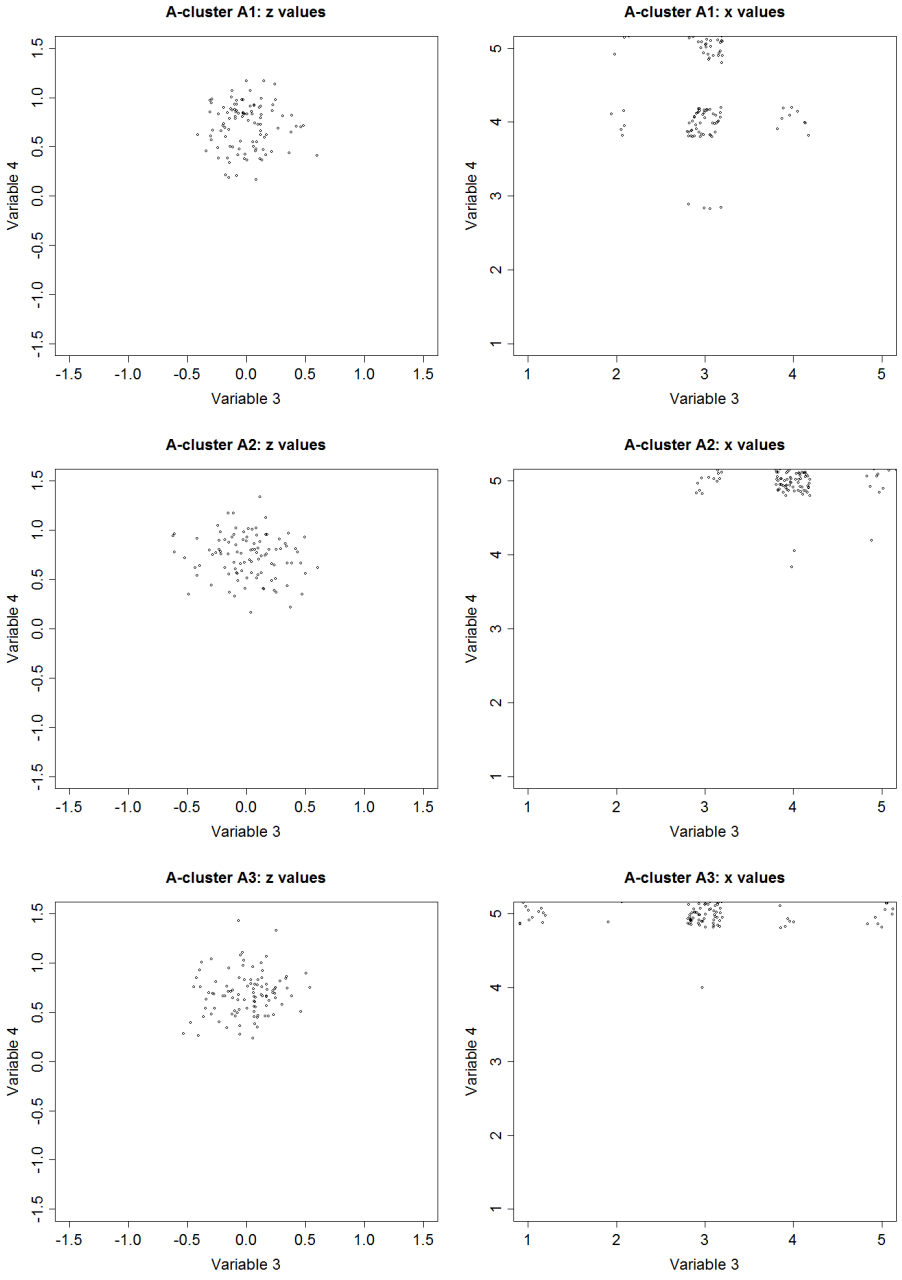


Figure 4.4: The effect of varying thresholds on categorised data. True latent scores for variables z_3 and z_4 in O-cluster O1 and three different A-clusters are shown on the left, whereas categorical answers resulting from applying appropriate A-cluster thresholds on z_3 and z_4 are presented on the right. Points in the plots on the right have been jittered to visualise the number of points in each category.

4.2 Latent model-based O- & hierarchical A- clustering

Due to a large computational burden connected with latent model-based clustering, only limited testing was possible. I limited my computations to five iterations of Algorithm 1, with five iterations of the EM algorithm within each. Despite these severe limitations, the computations took more than a week on a state-of-the-art 4-core machine. For the same reason I only used induced probabilities with the ARS measure for A-clustering. For a discussion regarding alternative ways of A-clustering see Section 4.3.2. To compensate for very low numbers of iterations, I started algorithm with reasonable starting values for O-cluster means, i.e. far from each other and near the true values. Earlier experiments with latent model-based clustering suggest that the algorithm finds good estimates, even if initialised with poor starting values. However, the convergence in such cases may be very slow.

Values of both fit criteria are presented in Figure 4.5. It shows that the values of both criteria increase monotonically, except for the initial iteration. However, the five iterations long run is too short to speculate about monotonic properties of the algorithm. In latent model-based clustering, due to using the UBN approach instead of full information maximum likelihood, both fit criteria have different scales, which makes choosing the best iteration more difficult than in clustering of expected values presented in the next section. Fortunately, because of the monotonic increase, the choice of the optimal iteration is simple and the fit criteria unambiguously suggest the last iteration.

Figure 4.6 illustrates the four true simulated mixture components (without considering the thresholds) and the true Γ-matrix of γ_{io} values. Figures 4.7 and 4.8 present the estimates of O-clusters after respectively 2^{nd} and 5^{th} iteration. These figures show that well separated clusters O2 and O3 ("red" and "green") are easy to identify in the algorithm, but lying close to each other clusters O1 and O4 pose some difficulties. As can be seen in Table 4.3, estimates of the means lie within the range of 0.3 - 0.4 from the true values, but also much greater differences are possible (variables 7 - 9 in the cluster O1). Figures 4.9 and 4.10 show that the thresholds estimates, in general, identify response patterns in the data correctly, but these estimates are not very precise and do not change substantially during the course of the algorithm. Tables 4.4 and 4.5 confirm the above conclusions. The O-clusters O2 and O3 are nearly perfectly identified, whereas the classification rate for the O-clusters O1 and O4 amounts to about $^2/_3$. As far as A-clusters are concerned, the algorithm identifies two clusters representing response styles quite well. The A-cluster A1 is also identified, but the identification in here is not so sharp. The percentage of correctly classified respondents in case of O-clusters amounts to 82.6% and in case of A-clusters to 76.2%.

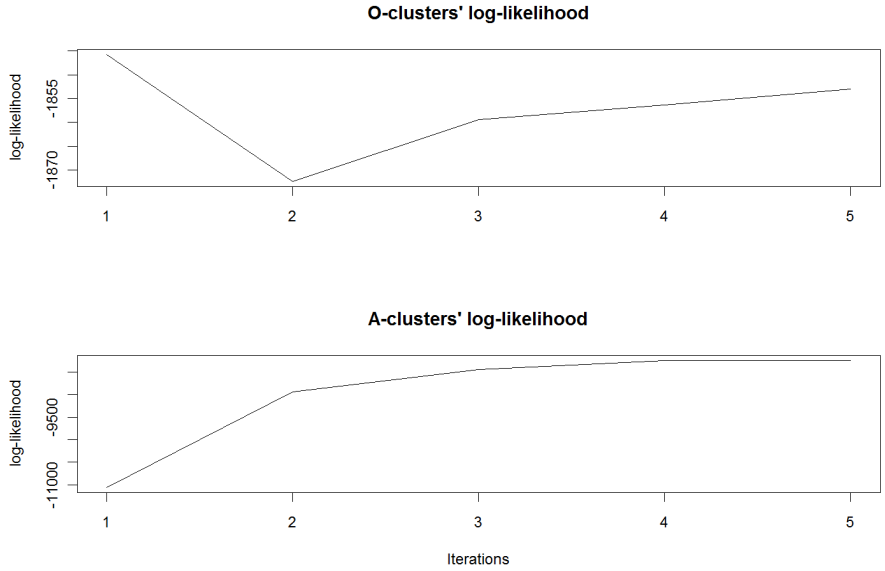


Figure 4.5: Values of the fit criteria for both sorts of clustering during the algorithm's course.

O-cluster	Mean vector μ_o^T								
O1	-0.12	-0.07	-0.17	0.88	0.88	0.90	1.56	2.84	1.91
O2	1.03	1.03	1.02	-0.90	-0.89	-0.92	0.30	0.31	0.33
O3	-0.93	-0.93	-0.91	1.03	1.05	1.03	-0.87	-0.90	-0.91
O4	-0.63	-0.65	-0.62	0.43	0.39	0.40	0.83	0.78	0.84

Table 4.3: Estimated means of four simulated O-clusters.

	estimated O-clusters			
true O-clusters	O1	O2	O3	O4
O1	193	0	4	103
O2	0	300	0	0
O3	0	0	298	2
O4	100	0	0	200

Table 4.4: True and estimated O-clusters after fifth iteration of the algorithm.

	estimated A-clusters		
true A-clusters	A1	A2	A3
A1	102	258	40
A2	334	66	0
A3	76	1	323

Table 4.5: True and estimated A-clusters after fifth iteration of the algorithm.

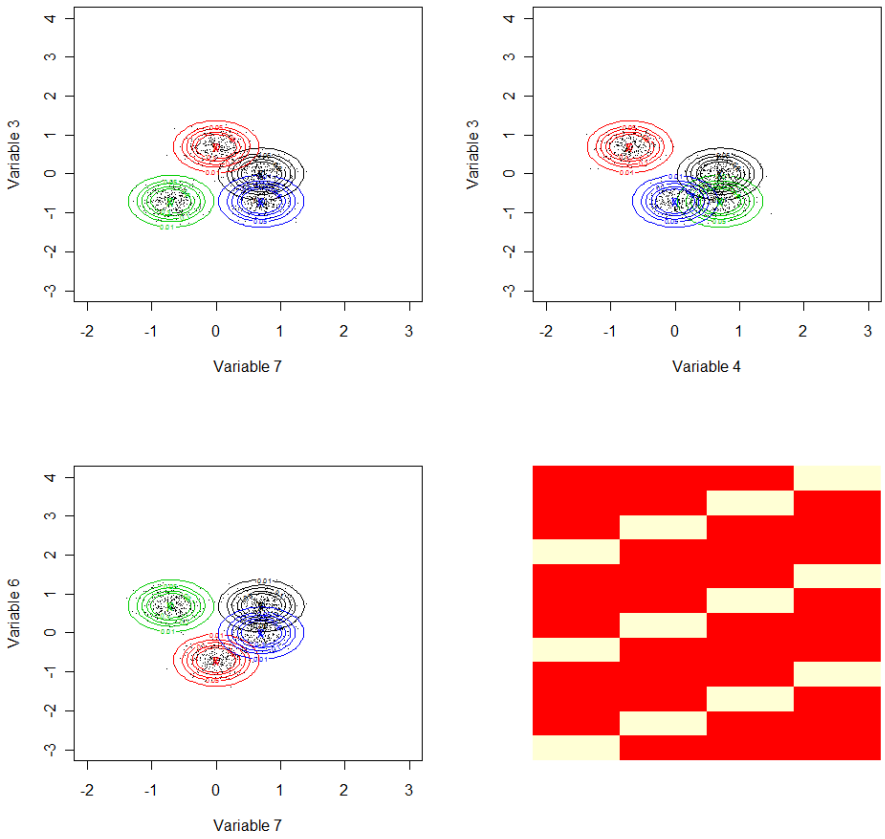


Figure 4.6: Three true bivariate marginal distributions with respect to variables 3, 4, 6 and 7 of the four mixture components and corresponding true matrix Γ of γ_{io} values. The Γ-matrix presented in the bottom-right panel consists of 1200 rows representing observations and 4 columns representing A-clusters. Colours ranging from red to light yellow represent values from 0 to 1 respectively. Since the true Γ-matrix consists of 0's and 1's only, only these two colours are present in the picture.

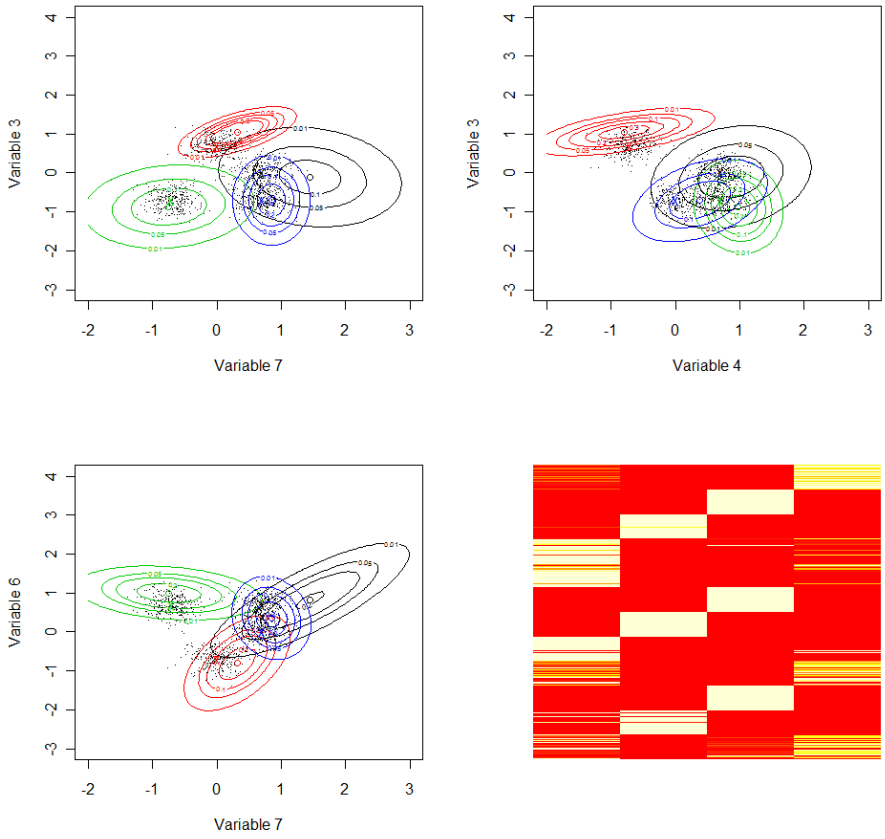


Figure 4.7: Three bivariate marginal distributions of four estimated mixture components and the Γ-matrix after second iteration.

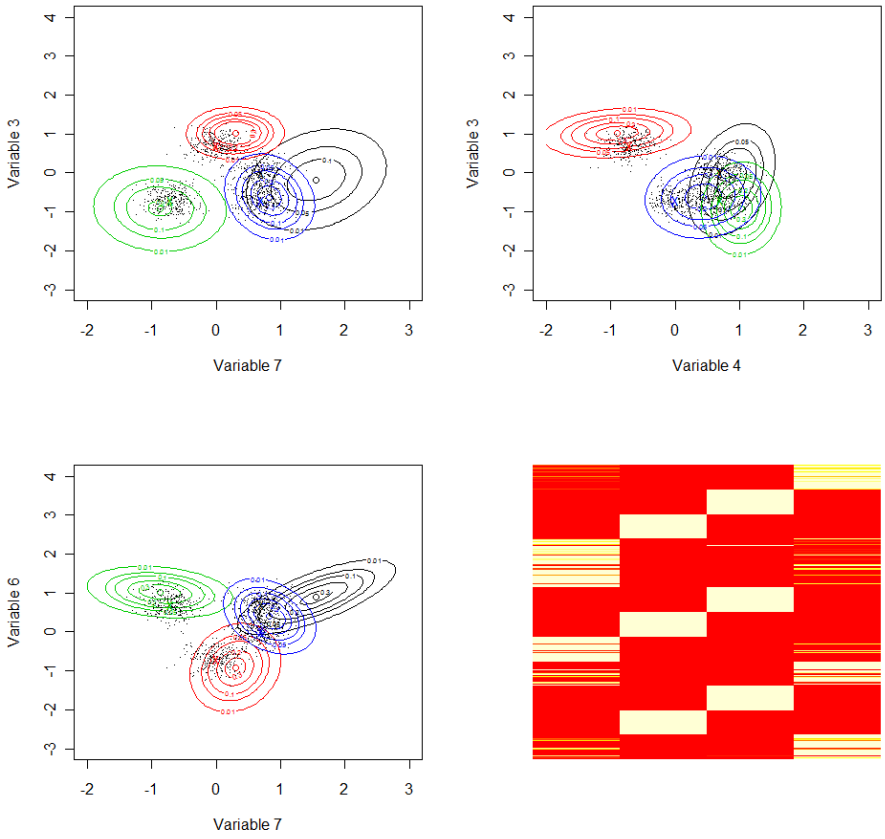


Figure 4.8: Three bivariate marginal distributions of four estimated mixture components and the Γ-matrix after fifth iteration.

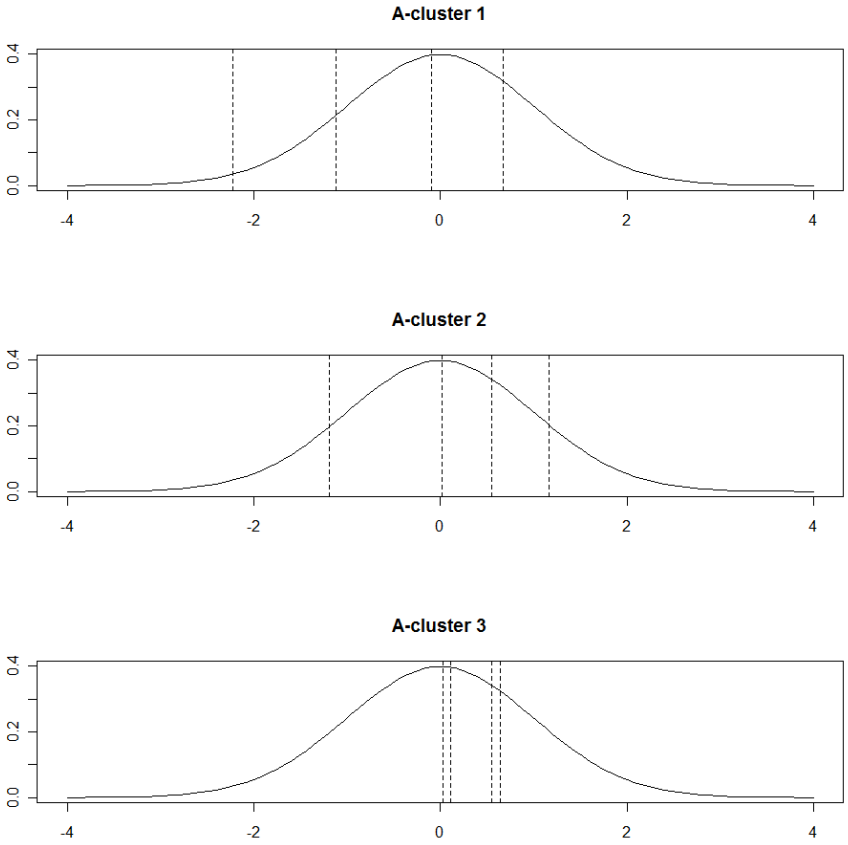


Figure 4.9: Thresholds estimates after second iteration.

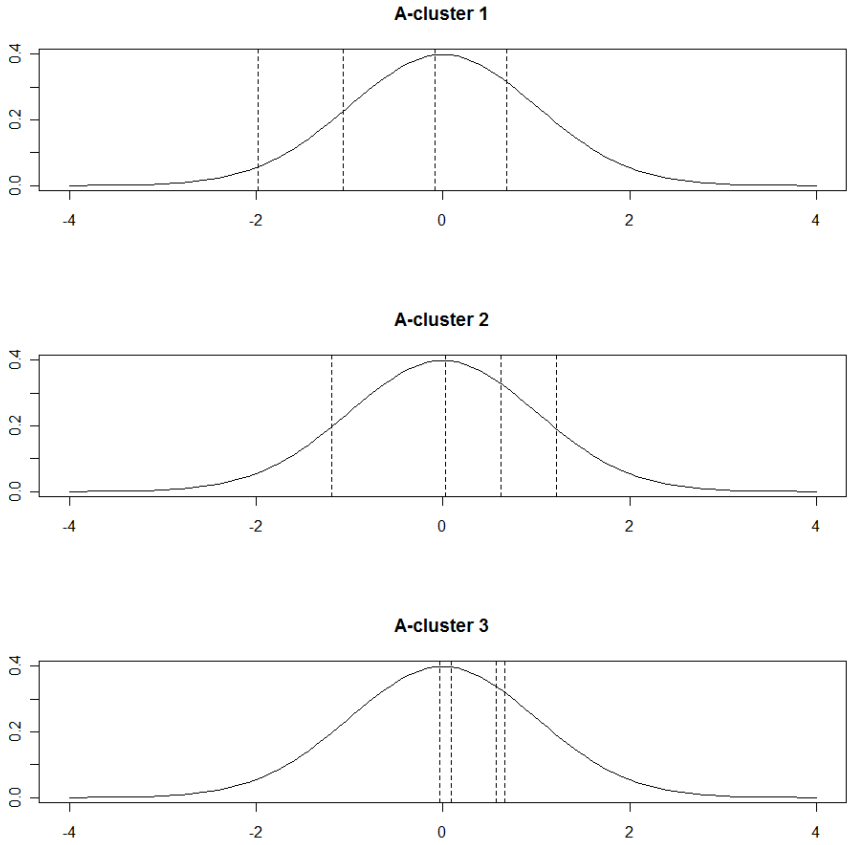


Figure 4.10: Thresholds estimates after fifth iteration.

4.3 Clustering using expected values of latent variables

I tested two variants of Algorithm 1 using expected values of latent variables. In the first one, I used hierarchical clustering for both sorts of clustering. In the second one, I used k-means clustering for the O-clusters and hierarchical clustering for A-clusters, similarly to the procedure described in Section 3.2. In both variants I also examine various approaches to A-clustering described in step 3 in Section 2.2.1.

4.3.1 Hierarchical O- & A- clustering

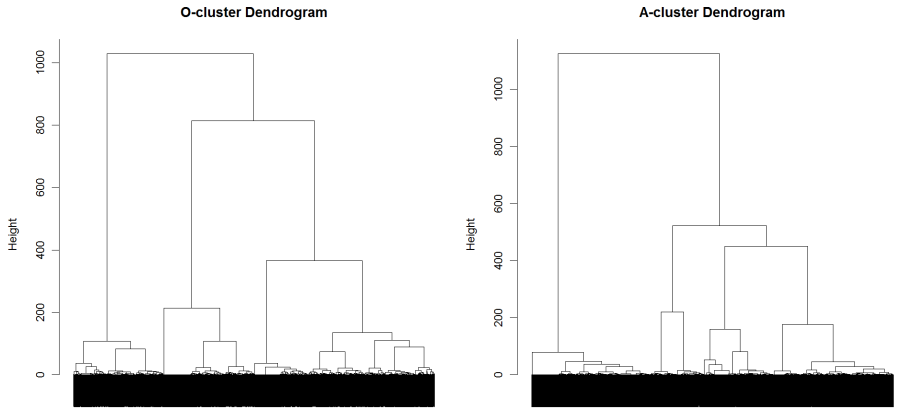
In clustering using expected values, initial partitions with respect to both sorts of clusters are required. Following recommendations given in Section 3.2 I start with classifying all observations into single O- and A- clusters. After that, I proceed with hierarchical O- and A- clustering, each time using Ward method with Euclidean metric and choosing the true number of clusters. Apart from Ward method, also single-, complete- and average- linkage as well as centroid methods were examined, but none of them revealed better performance.

Figure 4.11a presents the dendrogram of the first iteration of O-clustering. It suggests three rather than actual four clusters. A contingency table of the true and estimated O-clusters after selecting four clusters is presented in Table 4.6. We can see that, except for the O-cluster O1, all O-clusters are very well identified already in the first iteration. The analogous dendrogram and the contingency table for the first iteration of A-clustering are presented in Figure 4.11b and Table 4.7. Also here the dendrogram suggests a different to the true number of clusters. In this case two or four rather than three. The contingency table shows that the both A-clusters representing response patterns are relatively easy to identify. However, respondents from the true A-cluster A1 representing "normal" respondents have been distributed among all three estimated A-clusters.

	estimated O-clusters			
true O-clusters	O1	O2	O3	O4
O1	115	9	176	0
O2	1	0	0	299
O3	7	293	0	0
O4	297	0	3	0

Table 4.6: True and estimated O-clusters after first iteration of the algorithm.

The algorithm ran 40 iterations, after which both fit criteria defined in (2.36) and (2.40) as well the parameters' values practically do not change any more. The values of the fit criteria for all 40 iterations are presented in Figure 4.12. We see that after a few initial oscillations in the first few iterations, the algorithm achieves an equilibrium and converges gradually to final values. However, as shown later on, this is rather untypical behaviour for the analysed algorithm. In most cases the algorithm changes both partitions during its course, which



(a) (b)

Figure 4.11: Dendrograms of first iteration of the algorithm.

	estimated A-clusters		
true A-clusters	A1	A2	A3
A1	159	177	64
A2	394	1	5
A3	76	0	324

Table 4.7: True and estimated A-clusters after first iteration of the algorithm.

may cause substantial fluctuations of the fit statistics. Another characteristic of the algorithm is that even if convergence is achieved, it does necessarily indicate the optimum. This is the case in this example: despite the fact that in its course the algorithm converges to some value, the maximal value of the sum of both statistics is achieved already in the second iteration, before the plateau is achieved.

Figure 4.13 shows dendrograms for the second iteration of the algorithm, whereas Tables 4.8 and 4.9 present corresponding contingency tables. Both dendrograms give a stronger indication for the true number of clusters than their equivalents from the first iteration, especially as far as O-clusters are concerned. Of course, manual imposing of the correct number of clusters in the first iteration may have a substantial influence. However, the contingency tables do not indicate any improvement in the classification. In case of A-clusters, even some deterioration can be noticed as members of the true A-cluster A2 are more evenly distributed between estimated A-clusters A1 and A2. An examination of the final thresholds of the three estimated A-clusters in Figure 4.14 shows that the A-clusters A1 and A3 resemble relatively well the true thresholds. As for A2, however, we see a typical for the clustering of thresholds effect of setting outer thresholds to extremes, so that the probabilities for the corresponding categories are practically equal to 0, and respondents who have not chosen any

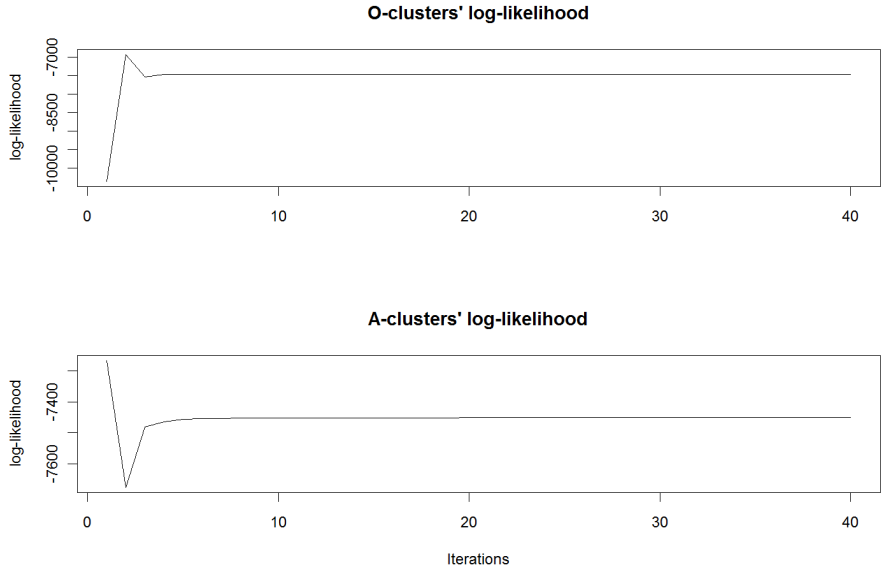
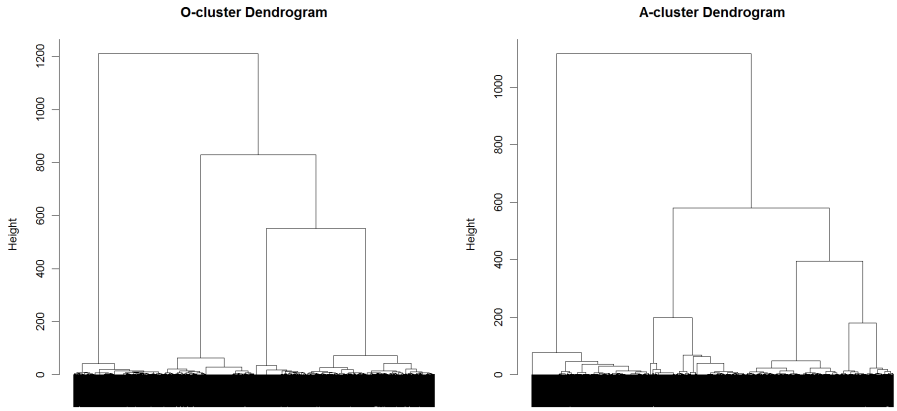


Figure 4.12: Values of the fit criteria for both sorts of clustering during the algorithm's course.

of the two lowest categories are clustered together. A comparison of the distributions of answers in the estimated (Figure 4.15) and the true (Figure 4.3) A-clusters confirm this conclusion.

	estimated O-clusters			
true O-clusters	O1	O2	O3	O4
O1	122	9	169	0
O2	0	0	0	300
O3	1	299	0	0
O4	299	0	1	0

Table 4.8: True and estimated O-clusters after iteration with the best fit.



(a) (b)

Figure 4.13: Dendrograms of iteration with the best fit.

	estimated A-clusters		
true A-clusters	A1	A2	A3
A1	261	76	63
A2	223	174	3
A3	6	70	324

Table 4.9: True and estimated A-clusters after iteration with the best fit.

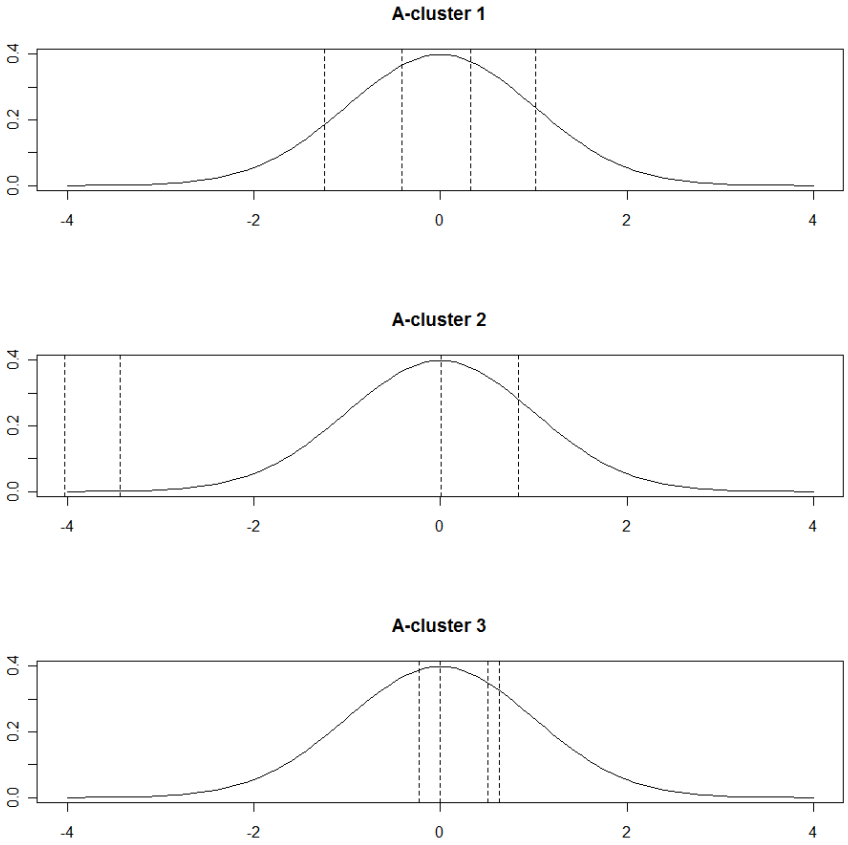


Figure 4.14: Thresholds of iteration with the best fit.

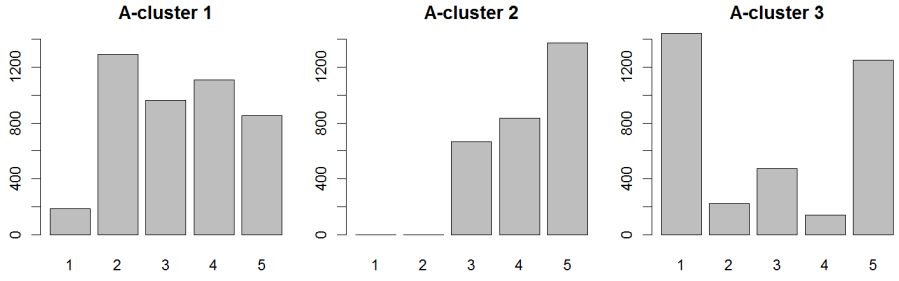


Figure 4.15: Distributions of answers within A-clusters in iteration with the best fit.

4.3.2 Effect of using alternative measures for A-clustering

A-clustering using induced probabilities

A direct use of individual thresholds in A-clustering results in a very strong tendency to bind together respondents who did not choose some of the extreme categories, which was also visible in the results presented in the previous section. This effect is so strong that it often dominates all other possible response patterns.

One way to overcome this shortcoming is to replace the thresholds with probabilities. First, individual thresholds are estimated. Next, they are applied to the standard normal distribution and probabilities between the thresholds, hereinafter "induced probabilities", are computed. These probabilities are clustered instead of thresholds. The advantage of using probabilities is that they are much more stable than the thresholds' values, firstly because they sum to 1 and secondly because the lack of responses in extreme categories does not result in extreme values of corresponding probabilities.

Similarly to the previous section, I iterated the algorithm 40 times. Both fit measures for this approach are presented in Figure 4.16. The maximum value for both criteria is achieved already in the second iteration, so I choose the results of this iteration as final. Figure 4.17 presents both final dendrograms. Whereas the O-cluster dendrogram indicates quite clearly the true number of clusters, the A-cluster dendrogram may indicate any number of clusters between 2 and 4. Tables 4.10 and 4.11 present contingency tables of the true and the estimated clusters. Comparing these tables with their counterparts from Section 4.3.1 reveals that whereas there is practically no change in the quality of the O-clusters, the A-clusters are better identified. Using probabilities instead of thresholds leads to a situation where in the A-cluster A1 there are 344 correctly classified cases compared to 261, in the A-cluster A2 there are 182 compared to 174; only in the A-cluster A3 there is a slight decrease from 324 to 318 correctly classified cases. All in all, replacing thresholds with induced probabilities increased the classification rate on the examined dataset from 59.5% to 70.3%. The barplots presented in Figure 4.18 reveal two characteristics which are typical for A-clusters obtained through clustering of induced probabilities. The first one is that they are not so sensible to the usage of extreme categories by the respondents as the A-clusters obtained through clustering of thresholds. All the A-clusters presented in Figure 4.18 contain answers in all categories. Even in the A-cluster A2 there are four answers in the first category. Unfortunately, the second characteristic of these clusters is that they tend to be difficult to interpret. The A-cluster A2 can be used again as a good illustration of the problem. Although it generally resembles the true A-cluster A2, the peaks in category 3 and 5 combined with a relatively low fraction of category 4 is difficult to explain.

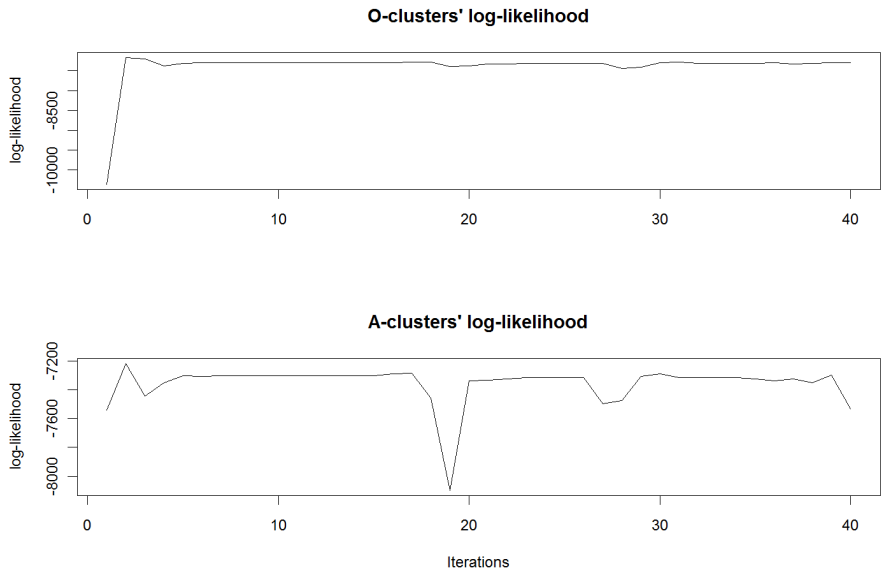


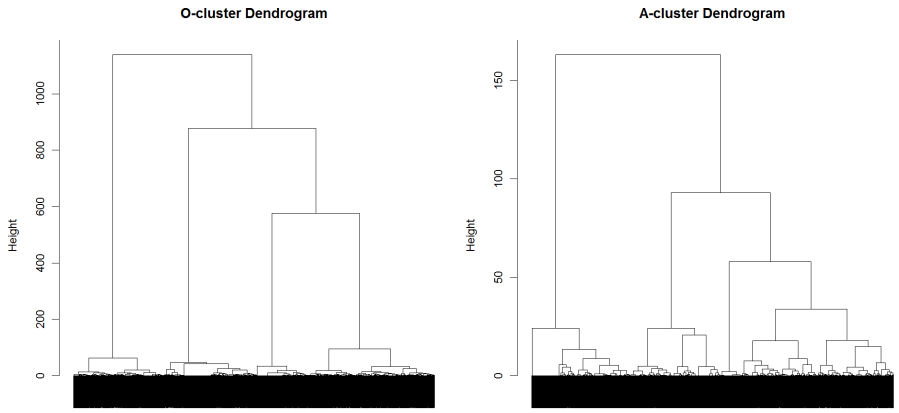
Figure 4.16: Values of fit criteria for both sorts of clustering during the algorithm's course.

	estimated O-clusters			
true O-clusters	O1	O2	O3	O4
O1	121	10	169	0
O2	1	0	0	299
O3	0	300	0	0
O4	299	0	1	0

Table 4.10: True and estimated O-clusters after iteration with the best fit.

	estimated A-clusters		
true A-clusters	A1	A2	A3
A1	344	49	7
A2	215	182	3
A3	14	68	318

Table 4.11: True and estimated A-clusters after iteration with the the best fit.



(a) (b)

Figure 4.17: Dendrograms of iteration with the best fit.

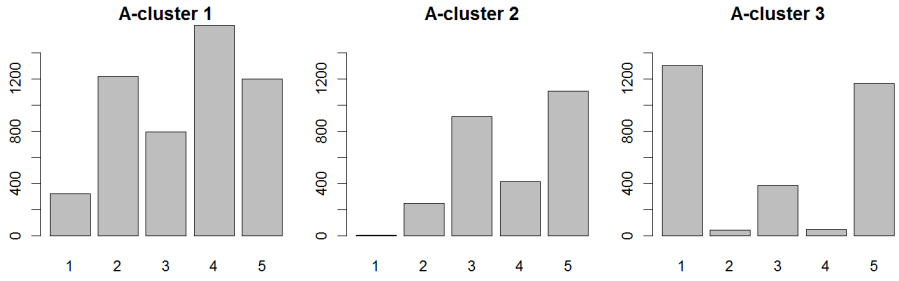


Figure 4.18: Distributions of answers within three estimated A-clusters.

A-clustering using induced probabilities and the ARS measure

One problem of A-clustering that remains unsolved after changing from the direct clustering of thresholds to the clustering of induced probabilities is the difficulty in identifying the clusters which represent the acquiescence response style. I address this problem by an explicit inclusion of an ARS measure in the set of variables used for clustering. The construction of the ARS measure is presented in step 3 in Section 2.2.1.

Since this measure has a different scale than the probabilities, and the scale of a variable affects the strength with which this variable influences clustering, I use a scaling factor for it. To find the optimal value for the scaling factor I examined various values in the range from 0 to 1.5.

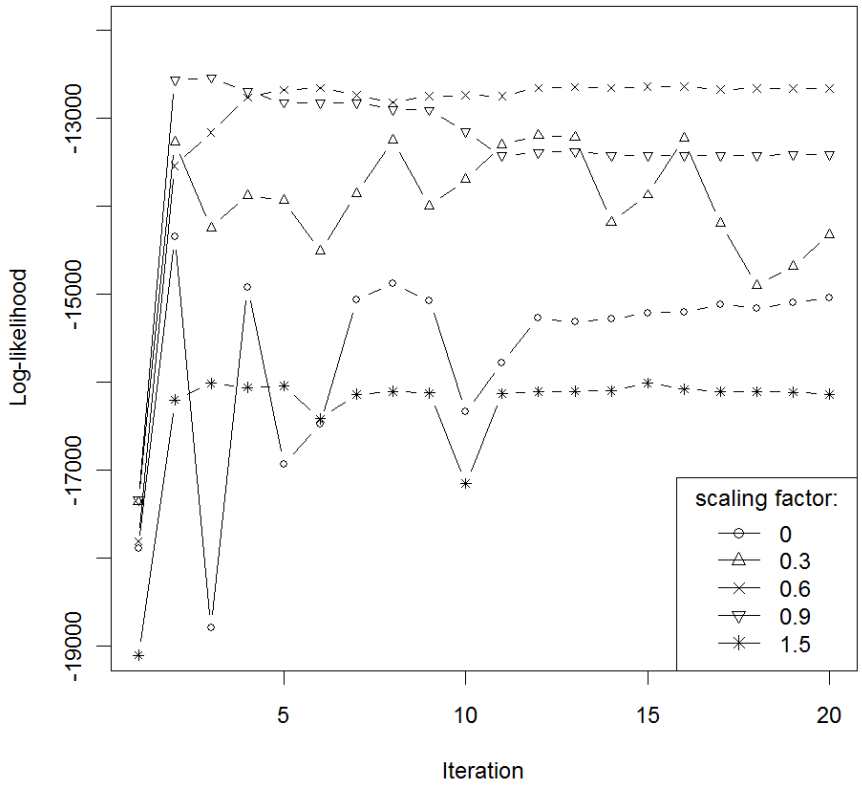


Figure 4.19: Effect of using different values of the scaling factor on A-clustering results.

Figure 4.19 presents the sums of the log-likelihoods defined in (2.36) and (2.40) for 20 iterations long runs of Algorithm 1. Sixteen values differing by 0.1 were tried. The figure presents selected results only. Figure 4.20 shows percentages of correctly classified cases for corresponding runs.

First of all, one may notice a great level of agreement between the log-

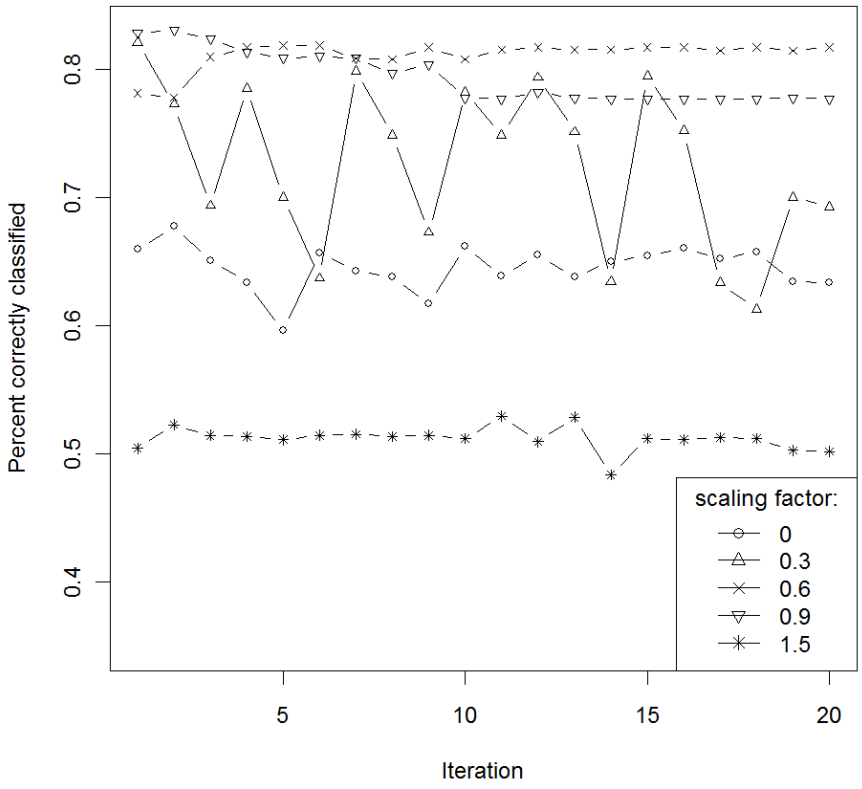


Figure 4.20: Effect of using different values of the scaling factor on A-clustering results.

likelihood values and the classification rates, which shows that the former is a good measure of assessing the model quality. The examination showed that the increasing of the scaling factor from 0 to about 0.5 increases the fit quality, values between 0.5 and 1.2 give the best results and that further increasing of the parameter value over 1.2 causes a decrease in model quality. The highest values of both measures were achieved using the scaling factor of 0.9. However, the fact that those values were achieved in the initial and possibly unstable phase of the algorithm and that after this phase both measures reveal a clear decline, casts some doubt on the stability of this result. Thus, I chose 0.6 as the optimal value of the scaling factor, for which the algorithm, after a few initial iterations, achieves a high, stable level with the maximum value only slightly lower than the maximum achieved with the scaling factor of 0.9.

After choosing the optimal value for the scaling factor, I examine the performance of A-clustering using induced probabilities and the ARS measure. Similarly to the previous cases, I run 40 iterations of the algorithm and choose the best iteration according to the fit criteria. Values of these criteria for all 40

iterations are presented in Figure 4.21. The sum of both criteria achieves its maximal value in the 16^{th} iteration and this one is chosen as the final. Figure 4.22 presents dendrograms for both sorts of clusterings. Like in other cases, the O-cluster dendrogram clearly indicates four O-clusters. The interpretation of the A-cluster dendrogram is much more vague, because it suggests two or three clusters. Tables 4.12 and 4.13 reveal that the classification with respect to market segments is practically identical to the previously investigated variants, whereas the classification of the response styles has substantially improved. In particular, the discrimination between the clusters A1 and A2 is much clearer. Including the ARS measure in the A-clustering process not only improved the rate of correctly classified respondents to 81.7%, but also made the interpretation of the obtained clusters much easier. Figure 4.23 shows that the resulting A-clusters resemble the true clusters quite exactly.

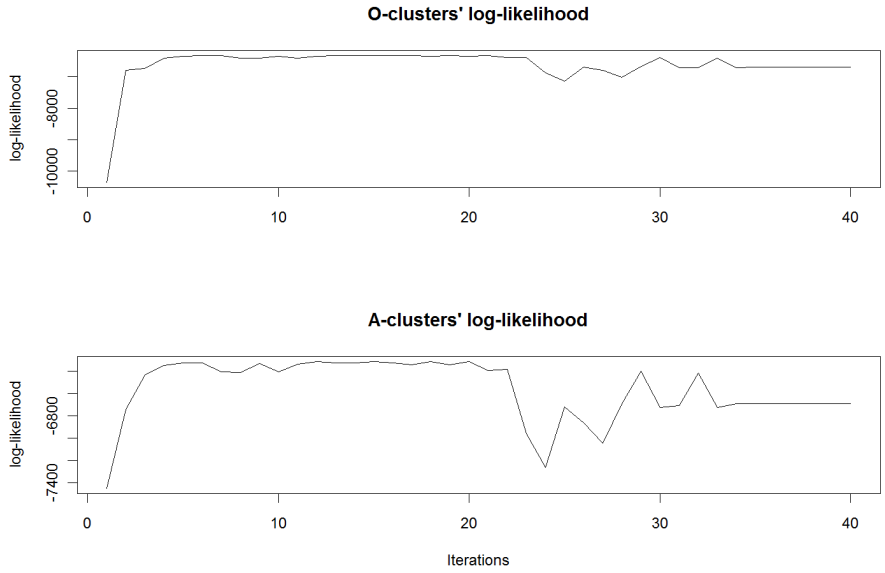
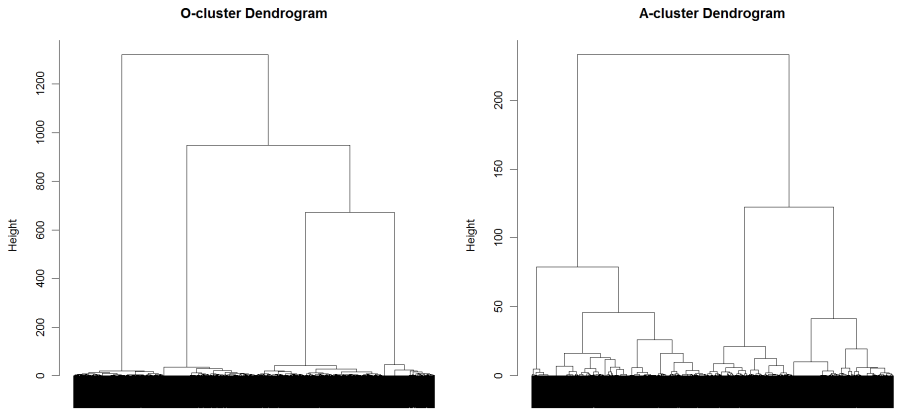


Figure 4.21: Values of fit criteria for both sorts of clustering, for A-clustering using induced probabilities and ARS measure.

	estimated O-clusters			
true O-clusters	O1	O2	O3	O4
O1	122	9	169	0
O2	1	0	0	299
O3	0	300	0	0
O4	299	0	1	0

Table 4.12: True and estimated O-clusters after iteration with the best fit.



(a) (b)

Figure 4.22: Dendrograms of iteration with the best fit.

	estimated A-clusters		
true A-clusters	A1	A2	A3
A1	112	272	16
A2	391	9	0
A3	77	6	317

Table 4.13: True and estimated A-clusters after iteration having the best fit.

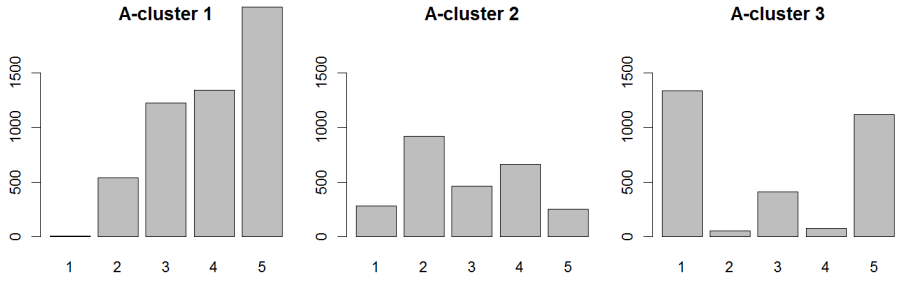


Figure 4.23: Distributions of answers within three estimated A-clusters.

4.3.3 K-means O- & hierarchical A- clustering

This section presents the results of applying expected values version of Algorithm 1 with k-means algorithm used for O-clustering and hierarchical A-clustering, i.e. in the same way as described in Section 3.2. In the two sections that follow, I combine this approach with two possible ways of A-clustering: (1) clustering of thresholds and (2) clustering of induced probabilities with the ARS measure.

A-clustering using thresholds

Similarly to Section 4.3.1 the algorithm is initialised by classifying all observations to single O- and A- clusters. After that, the main loop of the algorithm is iterated 40 times, and the true number of clusters is chosen in every iteration.

Tables 4.14 and 4.15 present the accuracy of both sorts of clustering after the first iteration. Here we can observe a similar situation to the one in Section 4.3.1, i.e. the O-clusters as well as the two A-clusters representing response styles are well identified already in the first step, but the respondents from the A-cluster A1 are distributed among other clusters.

	estimated O-clusters			
true O-clusters	O1	O2	O3	O4
O1	243	53	1	3
O2	0	0	300	0
O3	4	0	0	296
O4	14	286	0	0

Table 4.14: True and estimated O-clusters after first iteration of the algorithm.

	estimated A-clusters		
true A-clusters	A1	A2	A3
A1	75	148	177
A2	7	392	1
A3	329	71	0

Table 4.15: True and estimated A-clusters after first iteration of the algorithm.

Figure 4.24 depicts changes of the fit criteria for the whole course of the algorithm. It shows a possibly unstable behaviour of the algorithm quite clearly. After the initial oscillation, it seems to converge like in Section 4.3.1. However, gradual changes in expected values of $\mathbf{z}$ lead to a reclassification of the O-clusters in iteration 22. After that, the algorithm continues with higher values for both fit criteria. The iterations 29 to 40 exhibit another characteristic pattern of the algorithm's behaviour, i.e. oscillating between two partitions. Also similarly to Section 4.3.1 the maximal fit value is achieved before the final iteration, in this case in the 28^{th}.

An examination of Tables 4.16 and 4.17 shows that applying the algorithm does not lead to better identification of the true clusters. The final O- as well

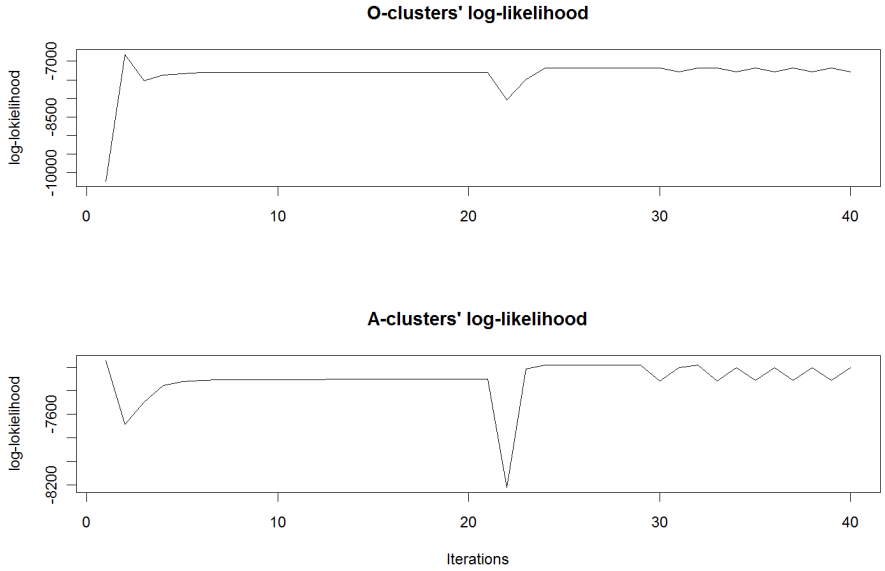


Figure 4.24: Values of fit criteria for both sorts of clustering during the algorithm's course.

	estimated O-clusters			
true O-clusters	O1	O2	O3	O4
O1	220	3	1	76
O2	0	0	300	0
O3	1	298	0	1
O4	59	0	0	241

Table 4.16: True and estimated O-clusters after iteration with the best fit.

as both A- partitions show poorer resemblance of the true partitions than their counterparts from the first iteration. Discrepancy between the final O-partition found by k-means and the true O-partition is also stronger than in hierarchical O-clustering despite higher values of the fit criteria for both sorts of clusters.

	estimated A-clusters		
true A-clusters	A1	A2	A3
A1	261	76	63
A2	223	174	3
A3	6	70	324

Table 4.17: True and estimated A-clusters after the iteration with the best fit.

A-clustering using induced probabilities and the ARS measure

Due to the substantial improvement in the A-clustering that resulted from the replacement of the thresholds with the induced probabilities together with the ARS measure in Section 4.3.1 I decided to examine the effect of this change in combination with O-clustering using the k-means algorithm. In this section I present the performance of this combination on the distance of 40 iterations of Algorithm 1.

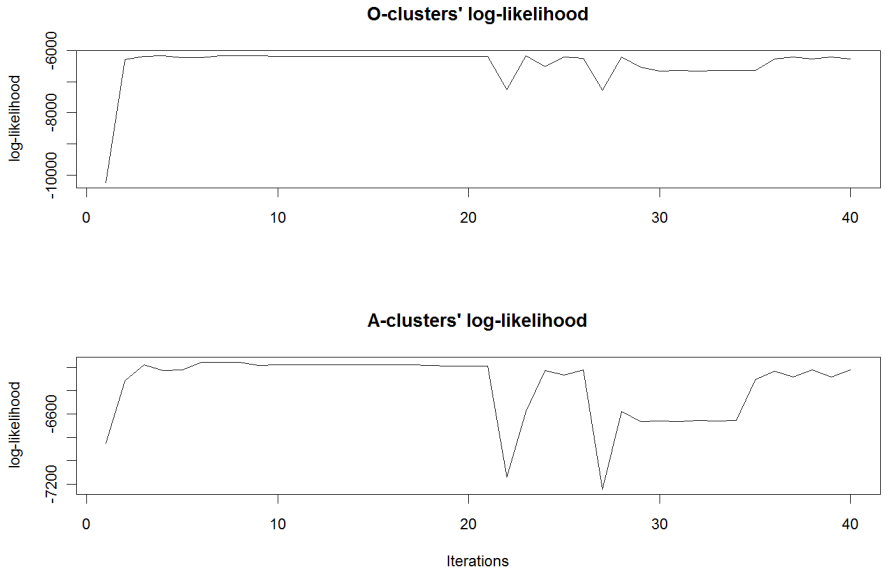


Figure 4.25: Values of fit criteria for both sorts of clustering.

	estimated O-clusters			
true O-clusters	O1	O2	O3	O4
O1	3	1	84	212
O2	0	300	0	0
O3	298	0	0	2
O4	0	0	292	8

Table 4.18: True and estimated O-clusters after iteration with the best fit.

The values of the fit criteria are depicted in Figure 4.25. As can be seen, having reached a relatively high plateau in the initial phase, the model fit decreases significantly after the 21st iteration. The sum of both fit criteria achieves its maximum in the 8th iteration. Despite the lack of improvement in later iterations, this combination outperforms all the previously analysed variants. The sum of both criteria in the 8th iteration amounts -12 312.99. This is better than the analogous A-clustering combined with hierarchical O-clustering, which achieved the fit equal to -12 638.64 and much better than the third ranked

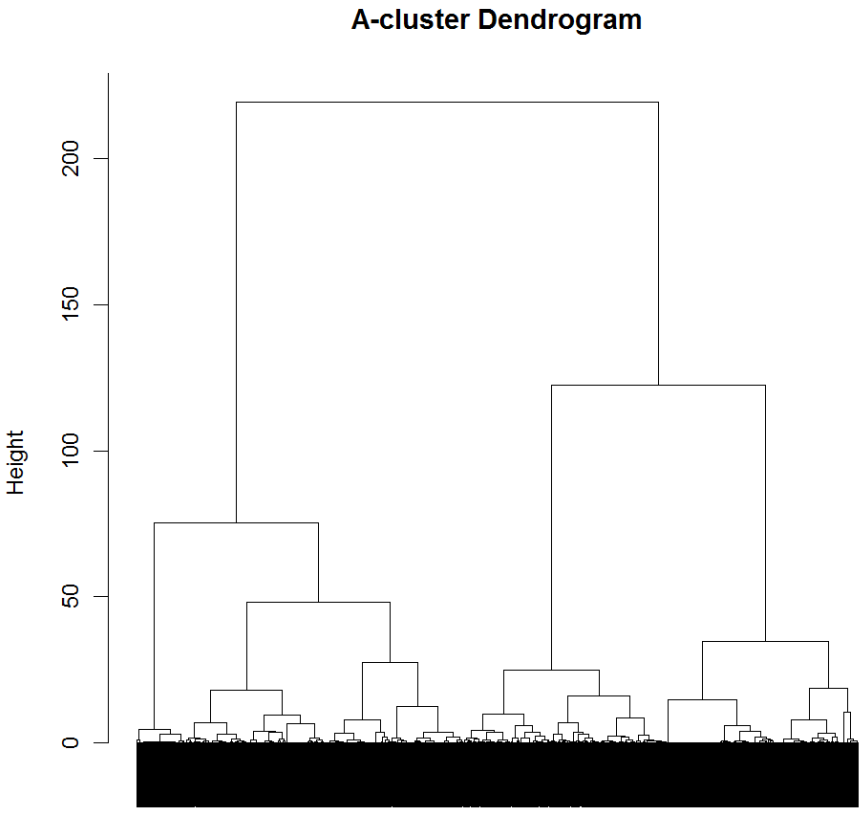


Figure 4.26: A-dendrogram of the optimal iteration.

variant of k-means O-clustering combined with A-clustering of thresholds (-14 355.23). The A-cluster dendrogram for the optimal iteration of this variant presented in Figure 4.26 is almost identical to its equivalent in Section 4.3.2 (A-clustering using induced probabilities and the ARS measure) and similarly suggests two or three A-clusters. Contingency tables presented in Tables 4.18 and 4.19 show that this variant identifies all O- and A- clusters correctly. In every O-cluster at least $^2/_3$ of cases are correctly classified and an analogous value for the A-clusters amounts $^3/_4$. The distributions of selected categories within the A-clusters presented in Figure 4.27 are very close to those obtained when this variant of A-clustering was combined with hierarchical O-clustering.

	estimated A-clusters		
true A-clusters	A1	A2	A3
A1	301	91	8
A2	14	383	3
A3	23	69	308

Table 4.19: True and estimated A-clusters after iteration with the best fit.

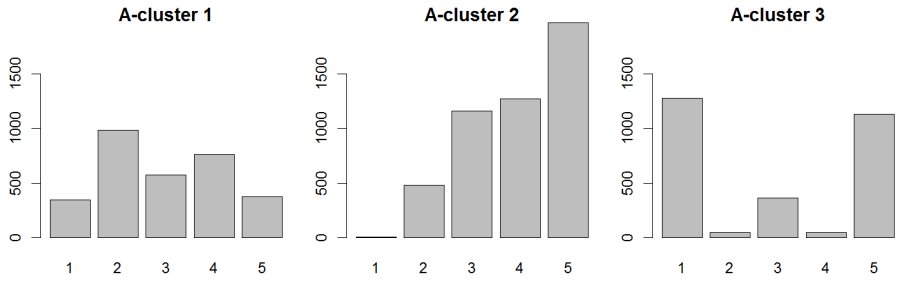


Figure 4.27: Distributions of answers in three estimated A-clusters.

4.4 Conclusions

Table 4.20 summarises the results obtained in this chapter. It presents all examined combinations of various variants of O- and A- clustering along with the fit statistics and classification rates under the limitations imposed in my simulations. We can clearly see that the great computational cost of latent model-based O-clustering supported with its appealing theoretical properties does not lead to high classification accuracy. Quite the opposite, the O-cluster classification rate of the latent model-based clustering is clearly the poorest. Furthermore, the classification rate with respect to response styles is the worst among the combinations using A-clustering of probabilities and the ARS measure (P+A A-clustering). Different definitions of the log-likelihoods for the latent model-based clustering and methods based on expected values make these measures impossible to compare between the two groups of algorithms. What is most striking among methods that use the expected values of the latent variables is a great difference between the variants using P+A A-clustering compared with other variants of A-clustering. Using P+A A-clustering improves both sorts of clustering by about 1000 points in terms of log-likelihoods. This gain in log-likelihood is mainly the consequence of much better A-clustering. The classification rates for A-clusters are even up to 20% better than in case of other variants of A-clustering. Although the results of O-clustering are quite stable for all examined variants even here, combined with k-means O-clustering the P+A variant of A-clustering brings some gain in the classification quality. Changing from the A-clustering of thresholds to probabilities does not seem to improve the log-likelihood, but the classification rate for A-clustering of probabilities is higher by about 7%. Finally, both variants which use O-clustering perform better than their hierarchical counterparts. However, the differences are rather small both in terms of log-likelihoods and classification rates. The only exception is the 3% difference in the O-clustering classification rates between the two variants using P+A A-clustering.

An important aspect one should keep in mind when comparing the three approaches to A-clustering, is the tendency to identify particular kinds of thresholds, which is not reflected in Table 4.20. The main problem of direct clustering of thresholds is its great sensitivity to the lack of answers in extreme categories. In such situation, the algorithm tends to cluster together respondents who did not choose extreme categories. According to professor Leisch's suggestion, in order to weaken this tendency I used induced probabilities. Clustering probabilities weakens substantially the above mentioned tendency. However, the resulting clusters may be difficult to interpret and it does not solve the other problem, which is difficulty in identifying respondents exhibiting acquiescence. This problem is addressed by explicit including an ARS measure among variables used in the clustering. This increases greatly the classification accuracy and results in well interpretable clusters.

O-clustering	A-clust.[1]	Total LL[2]	O-LL[3]	A-LL[4]	O-CR[5]	A-CR[6]
k-means	P+A	-12 312.99	-6 156.68	-6 156.31	91.83%	82.67%
hierarchical	P+A	-12 638.64	-6 313.54	-6 325.10	88.92%	81.67%
k-means	T	-14 355.23	-7 177.64	-7 177.59	88.25%	63.25%
hierarchical	P	-14 381.38	-7 164.60	-7 216.78	88.92%	70.33%
hierarchical	T	-14 590.46	-6 914.71	-7 675.75	88.67%	63.33%
latent model-based	P+A	-10 090.15[7]	-1 852.97[7]	-8 237.18[7]	82.60%	76.20%

Legend:
1. A-clustering: T - thresholds, P - probabilities, P+A - probabilities and the ARS measure.
2. Sum of the O-cluster and A-cluster log-likelihoods.
3. O-cluster log-likelihood.
4. A-cluster log-likelihood.
5. O-cluster classification rate.
6. A-cluster classification rate.
7. For latent model-based clustering different definitions of log-likelihood are used.

Table 4.20: Performance of various variants of O- and A- clustering.

Chapter 5

Real data example

This chapter presents an application of the methods presented in the thesis to real data. Due to a significant computational expense of the latent model-based clustering and the fact that the number of O-clusters is unknown, I present here only methods that are based on clustering expected values of latent variables, described in Section 2.2.2. For both types of clustering, I use hierarchical clustering.

5.1 Data

In my analysis I use the data on various attributes of fast food chains collected by Dolnicar et al. (2010). Respondents were asked to express their level of association between five fast food brands and ten attributes. The attributes are presented in Table 5.1. From the dataset described in (Dolnicar et al., 2010) I use seven-point scale responses only, which reduces the dataset to 715 respondents. For the purpose of the analysis, the data were recoded so that 7 means the strongest association, 1 the strongest dissociation and 4 is a neutral answer. In my analysis I limit O-clustering to two brands only: McDonald's and Subway. Using all brands could result in a large number of O-clusters, which would make the interpretation of the results difficult. The choice of the two

No.	Attribute
1	yummy
2	fattening
3	greasy
4	fast
5	cheap
6	tasty
7	healthy
8	disgusting
9	convenient
10	spicy

Table 5.1: Attributes examined in the survey.

brands mentioned above is motivated by the fact that they are likely to evoke very different associations as they target different fast food market segments. However, in A-clustering I use all the five brands to utilise the whole available information about the response styles.

Figure 5.1 presents distributions of answers for all examined attributes for both of the analysed brands. We see that McDonald's is usually associated with most of the attributes. The three exceptions are "healthy", "disgusting" and "spicy" which recall slightly negative association or no association at all. Subway is associated with the half of examined attributes, two attributes do not recall neither positive nor negative association and three attributes: "fattening", "greasy" and "disgusting", first two of which are particularly characteristic, are strongly dissociated with this brand. Despite these differences, marginal distributions of categorical answers within both brands are very similar and do not deviate from the distribution that was computed using all brands, which we can see in Figure 5.2. This figure also shows that positive categories are more often chosen than negative ones. This is to be expected given the information from Figure 5.1 and is probably the result of the questions' content rather than of the response style. However, the method presented in this thesis does not allow answering this question. Instead, it allows to identify clusters with respect to deviation from this "standard" response behaviour.

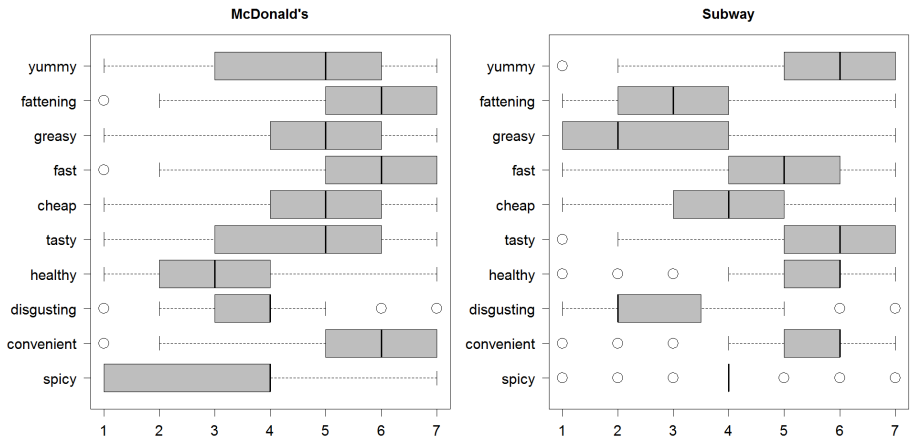


Figure 5.1: Distributions of categorical answers for McDonald's and Subway within examined attributes.

5.2 Analysis

It was a priori not clear how many O- and A- clusters there were to be expected in the analysed dataset. One way to determine the number of clusters is to apply hierarchical clustering and to use information from the dendrogram. However, analyses using simulated data suggest that in Algorithm 1 the structure of the dendrogram might be strongly influenced by the number of clusters chosen in the previous iteration. To avoid possible bias caused by this effect, I examined every potential combination of O- and A- clusters in a separate run.

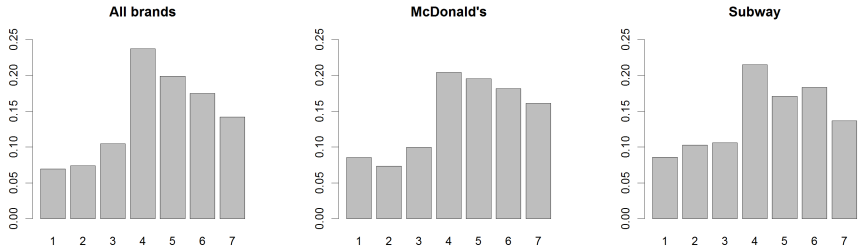


Figure 5.2: Marginal distributions of categorical answers in the whole dataset as well as in McDonald's and Subway subsets.

An initial exploratory analysis suggested that there were three O-clusters and no more than ten A-clusters. Thus, I ran twenty iterations of Algorithm 1 for every fixed combination of two to five O-clusters and two to ten A-clusters. In every combination, I used induced probabilities with the ARS measure for A-clustering. This gave thirty-six possible combinations.

To choose the optimal number of clusters various criteria were examined. Maximal log-likelihoods achieved for every of the thirty-six combinations are presented in Figure 5.3. It is little surprise that almost all of the log-likelihoods increase with the increasing number of both O- and A- clusters. Because of that I adapt to the proposed method classical model selection criteria AIC and BIC. Contrary to usual models, in the analysed method there are two sorts of log-likelihoods, separately for O- and A- clusters. To account for both of them, I use their sum instead of the standard $2\ln(L)$ term. This leads to the following formulae:

$$pAIC = -\ln(L^{OA}) + \quad 2 \times npar, \tag{5.1}$$

$$pBIC = -\ln(L^{OA}) + \ln(N) \times npar, \tag{5.2}$$

where $\ln(L^{OA})$ is the sum of O- and A- log-likelihoods and $npar$ equals $2 \times J \times O + (K-1) \times A$. It is important keep in mind that in this example J equals 50 and not 20. Although only 20 variables are used for O-clustering, means and variances must be estimated for all 50, because we need them for estimating optimal thresholds. Since the above formulae differ from the original AIC and BIC definitions I refer to them as pseudo- AIC/BIC and denote $pAIC$ and $pBIC$ respectively.

Figure 5.4 presents the values of the both criteria for all estimated models. They both suggest a model with five O-clusters and nine A-clusters. However, it is likely that further increase of the number of clusters would lead to even lower values of these criteria. Unfortunately, the solutions with many clusters are difficult to interpret and do not seem to be useful in understanding differences between respondents, so I look for solutions with fewer clusters. Also an examination of the dendrograms did not give a clear answer how many clusters to choose. In this situation I decided to choose the combination which is best interpretable. In my opinion, the combination of three O-clusters and six A-clusters produces the most meaningful clusters, and this one I choose as final.

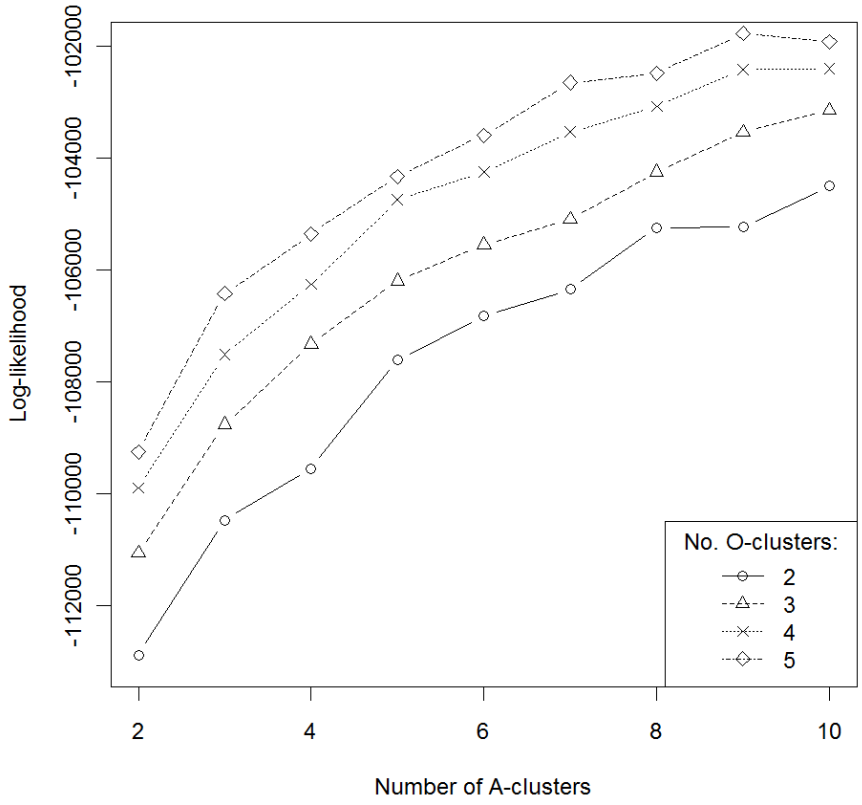


Figure 5.3: Maximal log-likelihoods achieved for every examined combination of O- and A- clusters.

Figure 5.5 shows distributions of the expected values of the latent variables in three estimated O-clusters. All the three clusters allow a good interpretation. The O-cluster O1 consists of average fast food customers. They share typical, moderately intense associations connected with both brands, which very closely resemble marginals shown in Figure 5.1. The O-cluster O2 contains Subway fans. In general they hold similar associations to the members of O-cluster O1, but their associations regarding Subway are more extreme. They clearly associate Subway with good taste ("yummy", "tasty") and healthy food (not "greasy", not "fattening", "healthy"). The answers for McDonald's in this cluster are much the same like those in the cluster O1, but also here we can see some signs of positive attitude toward fast food: rather positive answers for "yummy" and "tasty", a lack of negative association with "healthy". To conclude, this O-cluster may be summarised as grouping people with non-negative attitude toward fast food in general and particularly positive opinion about Subway. The O-cluster O3 exhibits the clearest pattern of the three O-clusters. Whereas the respondents in this cluster do not differ substantially from the

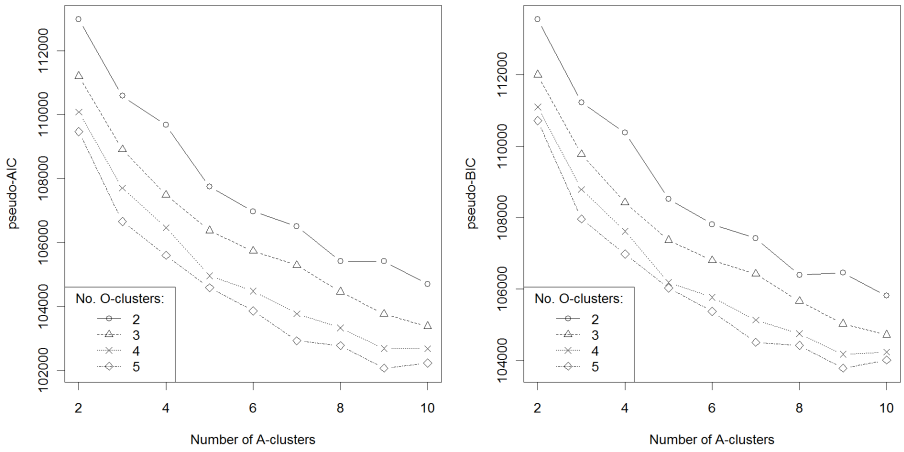


Figure 5.4: Pseudo- AIC and BIC values for examined combinations of O- and A- clusters.

population average with respect to their opinion about Subway, they are great McDonald's antagonists. They represent a very strong and homogeneous view that McDonald's is unhealthy ("greasy", "fattening", not "healthy") and does not taste good (not "yummy", not "tasty"). The strength of these negative associations is underlined by positive associations with "disgusting" (the only such case in all O-clusters).

Figure 5.6 presents the six estimated A-clusters. To describe them, I use the response styles listed by Baumgartner and Steenkamp (2001, Table 1). The A-cluster A3 resembles quite exactly the marginal distribution of answers in the whole dataset shown in Figure 5.2 and reflects the most typical response pattern for the whole population. I will use this A-cluster as a reference for comparisons with other A-clusters. Compared to the cluster A3, the clusters A2 and A4 represent narrow response range (RR). Both distributions exhibit a clear unimodal pattern where the middle categories are much more frequently chosen than the extreme ones. The difference between these clusters lies in acquiescence. Whereas in the cluster A4 mode of the distribution lies in the middle of the scale, in the cluster A2 it is shifted by one and lies at category 5. If the scale was balanced, it would mean that the cluster A2 exhibits acquiescence. However, in this case I only conclude that the respondents in the cluster A2 exhibit stronger acquiescence than those in the cluster A4. Even stronger concentration around the mode represent respondents in the cluster A5. This an example of midpoint responding (MPR). This response style in this survey is often result of the lack of knowledge of some of the rated brands. In such situations the respondents often choose middle categories for all attributes which results in the pattern presented in Figure 5.6. The two remaining A-clusters: A2 and A6 are examples of extreme response style (ERS). They both contain respondents who exhibit stronger than average tendency to use extreme categories. The barplots show that these are the only clusters where modes lie (also) at extreme categories. The difference between these two clusters is much the same as between the clusters A1 and A4, i.e. in the level of acquiescence. Whereas in the cluster A2 the distribution of answers is quite symmetrical, in

the A6 positive categories are clearly more often chosen than the negative ones. The closer look at the distribution of responses in the cluster A6 reveals that it resembles quite exactly the distribution in the A-cluster A3 with the only difference that the extreme categories are about twice as likely to be chosen. So, similarly as in case of the clusters A1 and A4 we have here two clusters, of which one is symmetric and the other left skewed.

Figure 5.7 shows the thresholds' estimates for all the A-clusters. A relatively symmetric distribution around 0 of most of those thresholds might suggest that the skewed distributions in Figure 5.6 are result of true respondents' opinions rather than of acquiescence. However, by the interpretation of the thresholds one must keep in mind how scale and location of the latent variables are fixed in the initialisation step (see description of the initialisation step in Section 2.2.1 for details). Initial mapping of the categories to the intervals defined by the quantiles of the standard normal distribution causes that always a distribution close to the population marginal will result in thresholds similar to the quantiles. This phenomenon is clearly visible in the Figure 5.8, where the most similar to the global marginals A-cluster A3 is compared to the standard normal quantiles. Because of that, unless any additional measures are used (see e.g. Baumgartner and Steenkamp, 2001, for proposals), any inference about acquiescence and response range in the estimated A-clusters should be relative to the population's marginal.

	A-clusters						
O-clusters	A1	A2	A3	A4	A5	A6	Σ
O1	87	77	97	47	17	41	366
O2	60	58	68	48	16	33	283
O3	7	21	12	5	2	19	66
Σ	154	156	177	100	35	93	715

Table 5.2: Cross-tabulation of estimated O- and A- clusters.

Finally, I examine the bivariate distribution of respondents between the estimated O- and A- clusters to check whether the O-clusters do not contain information about response styles or the A-clusters do not contain information about opinions. If that was the case, the both sorts of clusters should be highly dependent and form groups where certain O-clusters coincide only with certain A-clusters. This bivariate distribution is presented in Table 5.2. As we can see this distribution does not reveal any visible dependence. Every A-cluster is represented in every O-cluster. Although the χ^2 and the Fisher's tests of independence suggest rejecting the null hypothesis with p-values respectively 0.0016 and 0.0044[1], this dependence does not seem to threaten the validity of the results in a considerable manner. Consequently, I conclude that neither the O-clusters are distorted substantially by the response styles nor the A-clusters by the market segments.

[1] simulated p-value based on 10^5 replicates

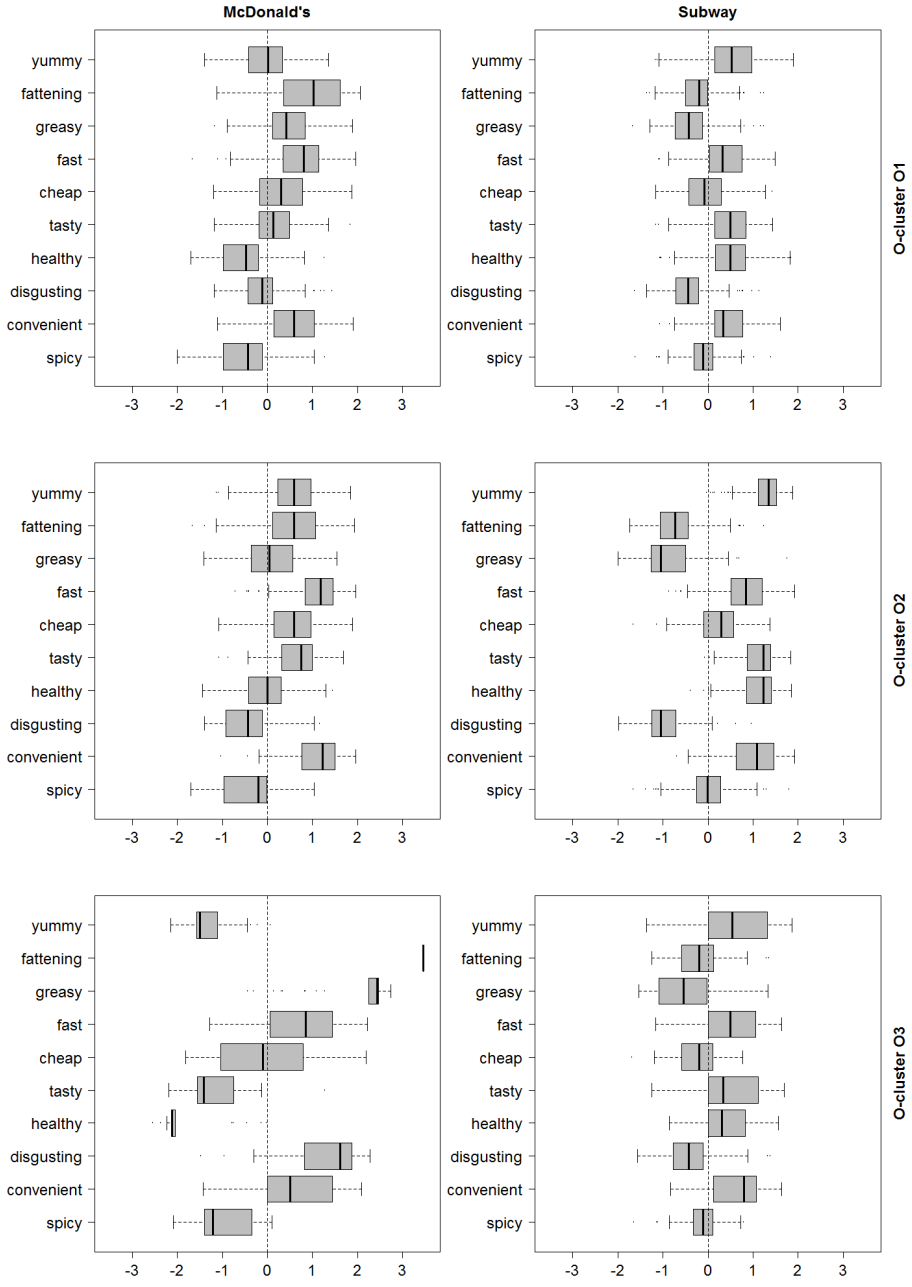


Figure 5.5: Three final O-clusters.

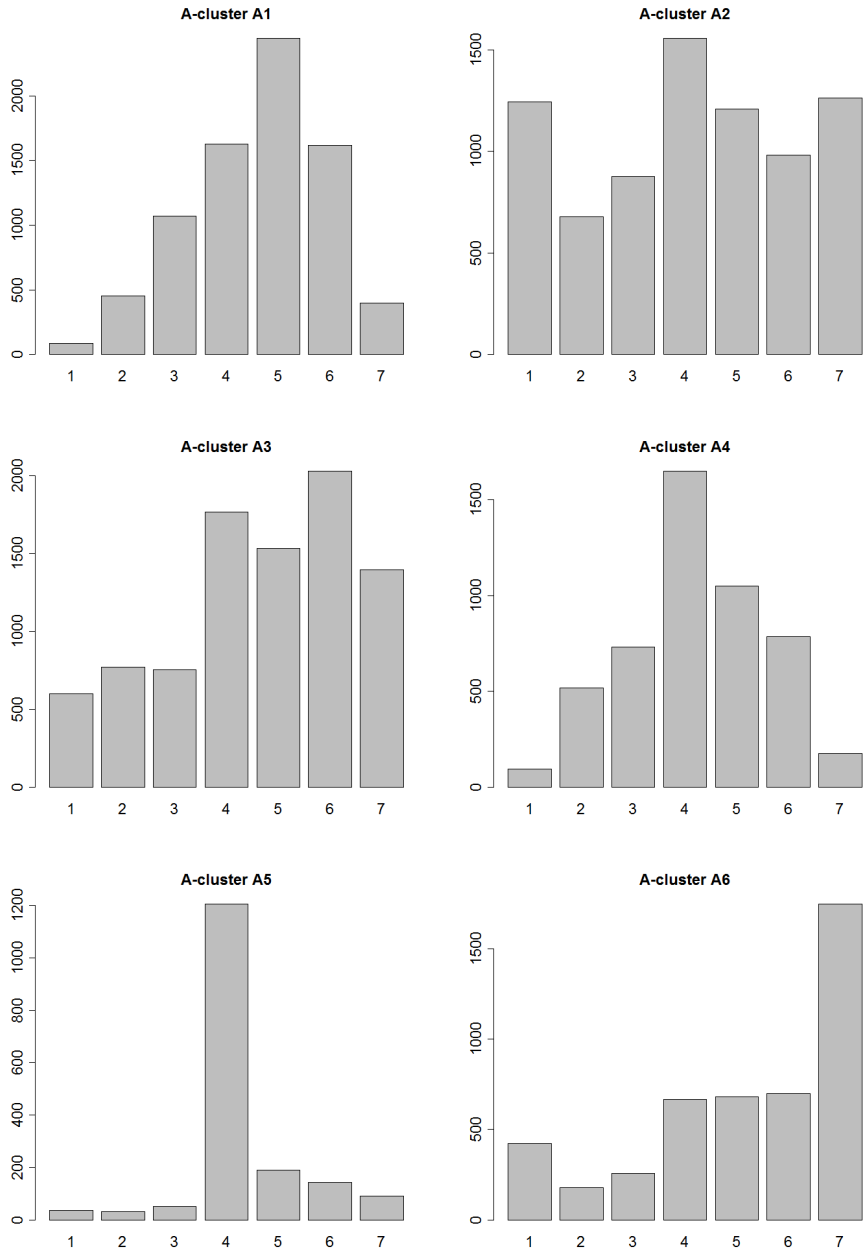


Figure 5.6: Six final A-clusters.

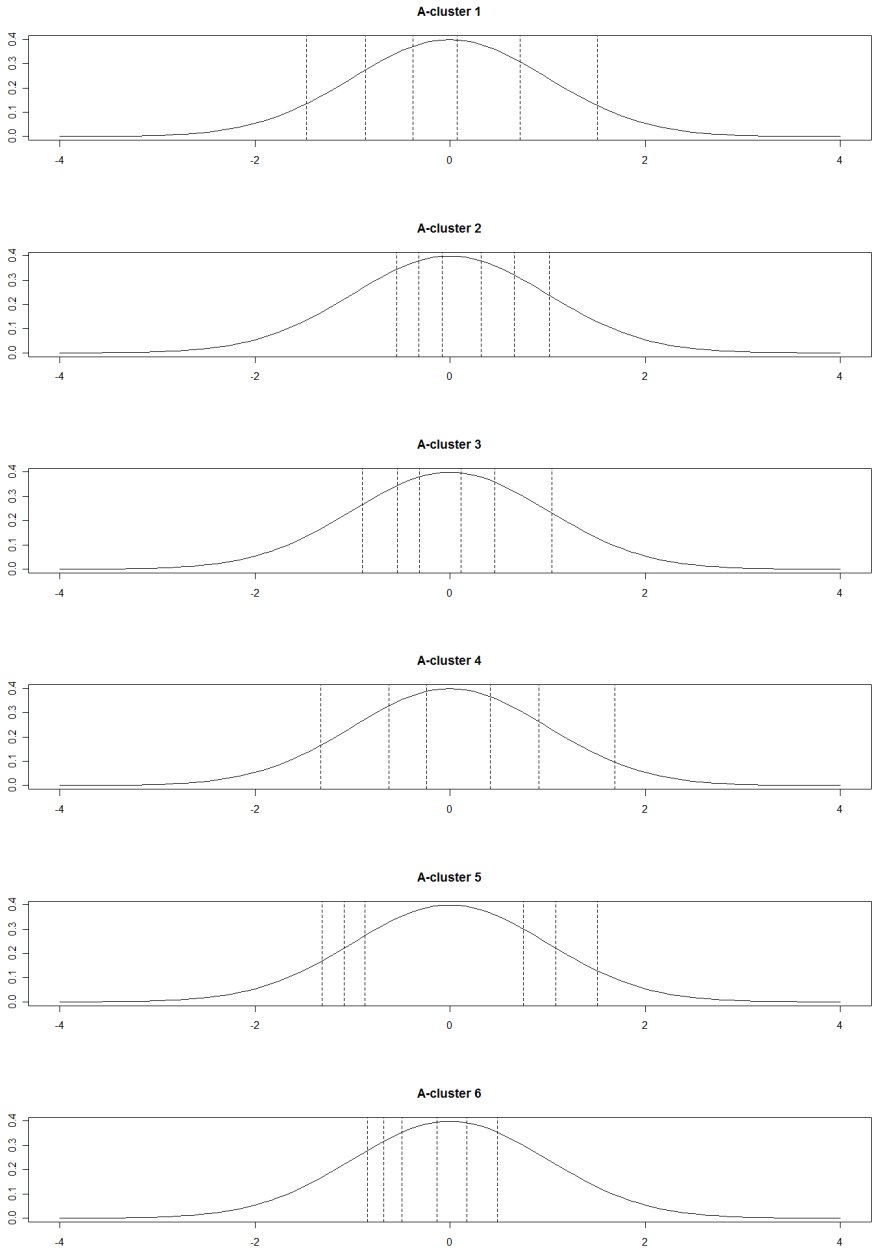


Figure 5.7: Thresholds' estimates for the final A-clusters.

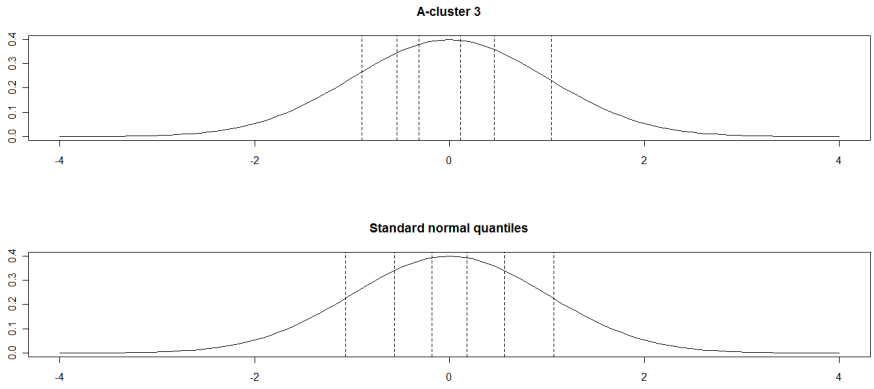


Figure 5.8: Comparison of thresholds' estimates for A-cluster A3 and quantiles of standard normal distribution.

Chapter 6

Discussion

In my thesis I deal with the problem of scale usage heterogeneity. Despite the vast existing literature describing this problem from psychological perspective there are only a few articles approaching this problem from statistical point of view and suggesting methods capable of accounting for heterogeneity in using rating scales. Instead of correcting for scale usage heterogeneity at an individual level, as it is commonly proposed, the method presented in the thesis consists in clustering respondents with respect to their response styles and use the clusters' estimates for accounting for the response style. The rationale for such an approach is an assumption that there is no need to treat the respondents' response styles individually. I assume that there are only several typical response styles which are exhibited by most of the respondents and cluster them trying to identify those styles. This allows, on the one hand, to reduce greatly number of parameters to estimate, on the other hand, to still have fixed, interpretable descriptions of the response styles (contrary to the random effects approach proposed by some authors) and to obtain a segmentation of respondents with respect to their response styles. I found in the literature only one article also proposing clustering with respect to response styles (Van Rosmalen et al., 2010). However, in this article the authors use the item response approach, contrary to the underlying latent variables approach used here.

To account for categorical character of data I use two different approaches. The first one - latent model-based clustering - seems to be a natural approach. Theoretically, it allows to estimate parameters of underlying latent multivariate mixture distribution without the need of estimating individual values of the underlying latent variables. However, both computation of the probabilities of hyperrectangular cut-offs of the multivariate normal distribution and the EM algorithm for estimating parameters of mixture distributions are computationally intensive and combination of both in the latent model-based clustering makes a computational burden connected with this method practically prohibitive, even with application of modern estimation methods, such as the "Underlying Bivariate Normal" approach. To overcome these computational problems I propose a second approach, which is based on computing and clustering expected values of the latent variables. This approach offers relatively easy to compute methods at the cost of some simplifying assumptions.

Another problem to solve was to find a method to cluster respondents with respect to their response style. Within the proposed framework, a natural way

to do this seems to be estimating thresholds for every respondent, and then, clustering them using any of the standard clustering techniques. However, this method proved to be very sensitive to usage of extreme categories, or strictly speaking, the lack of thereof. This practically dominated any other response patterns present in data. To solve this problem, following the suggestion of professor Leisch, I tried using probabilities induced by applying estimated thresholds to the standard normal distribution. This method solved the problem of sensitivity against using extreme categories. However, it still had problems with identifying acquiescence and often resulted in A-clusters that were difficult to interpret. To address also this problem, I explicitly included an ARS measure in A-clustering. The simulation results show that this improves greatly quality of estimated A-clusters and also results in A-clusters which are easier to interpret. Also real data application shows that this approach gives reasonable results.

The methods based on expected values of latent variables presented in the thesis gave reasonable results applied both to simulated and to real data. Nevertheless, there is still some potential for improvement and further research. Firstly, except for its computational requirements, latent model-based clustering is still a very interesting approach to O-clustering. Improvements in estimation methods for this approach could make it feasible. Two potential areas for improvement are developing faster estimation methods or finding an alternative distribution for the multivariate normal, which would have a closed-form integral. Another direction for further work is the reduction of simplifying assumptions used in expected values approach. The most important of them seems to be ignoring of correlations between latent variables. Finally, it is still not obvious what is the optimal way of clustering individuals with respect of response styles. Using induced probabilities with the ARS measure gives good results, but there still may be better approaches.

Appendix A

Explanation of the notation used in the thesis

Symbol	Explanation
a	Variable indexing A-clusters.
o	Variable indexing O-clusters.
i	Variable indexing individuals in the analyzed dataset.
j	Variable indexing questions in the analyzed dataset.
$k(j)$	the j-th element of the vector $\mathbf{k}$.
r	Variable indexing all possible response patterns (vectors $\mathbf{k}$). A "response pattern" is defined as a vector of categorical answers to all questions in the questionnaire.
n_{aor}	Number of response patterns r given by the members of the A-cluster a and the O-cluster o in the analyzed dataset.
n_o	Number of the members of the O-cluster o in the analyzed dataset.
$p_{a\|o}$	A fraction of members of the A-cluster a in the O-cluster o.
$p_{r\|ao}$	A fraction of answers r in the intersection of the A-cluster a and the O-cluster o.
$\dot{p}_{a\|o}$	A fraction of probabilities γ_{io} of the members of the A-cluster a in the O-cluster o.
$\dot{p}_{r\|ao}$	A fraction of probabilities γ_{io} with responses r in the intersection of the A-cluster a and the O-cluster o.
x_{ij}	Categorical answer of the individual i to the question j.
z_{ij}	Latent continuous variable representing the "real" attitude of the individual i to the question j.
$\mathbf{k}$	A vector of integers representing categorical answers to all questions in the questionnaire.
$\mathbf{x_i}$	A vector of categorical answers of the individual i to all questions in the questionnaire.
$\mathbf{z_i}$	A vector of the values of the latent attitudes of the individual i to all questions in the questionnaire.
A	Number of A-clusters.

Table A.1: Notation summary.

Symbol	Explanation
O	Number of O-clusters or a set of all O-clusters' indexes (depending on context).
J	Number of questions in the questionnaire.
K	Number of categorical answers.
N	Number of cases in the analyzed dataset.
μ_{oj}	Expected value of the latent attitude to the question j in the O-cluster o.
σ^2_{oj}	Variance of the the latent attitude to the question j in the O-cluster o.
ρ_{jl}	Correlation between variables j and l.
$\tau_{a,k}$	A threshold between the categories k and $k+1$ in the A-cluster a.
γ_{io}	Probability that the person i belongs to the O-cluster o (in latent model-based O-clustering).
λ_o	Proportion of the O-cluster o in the mixture density.
$\pi^{(j)}_{k\|ao}$	Probability of choosing category k in the question j by a person belonging to the A-cluster a and the O-cluster o.
$\pi^{(jl)}_{k,m\|ao}$	Probability of choosing category k in the question j and category m in the question l by a person belonging to the A-cluster a and the O-cluster o.
$\pi_{r\|ao}$	Probability of choosing the response pattern r by a person belonging to the A-cluster a and the O-cluster o.
μ_o	A vector of expected values of the latent attitudes to all questions in the questionnaire in the O-cluster o.
Γ	A matrix of γ_{io} values.
Σ_o	Covariance matrix of the latent attitudes to all questions in the questionnaire in the O-cluster o.
$\boldsymbol{\tau}_a$	A vector of thresholds in the A-cluster a.
$\boldsymbol{\tau}_i$	A vector of individual thresholds for the individual i.
$I()$	A set of respondents for which condition in brackets is true.
$I(a)$	A set of respondents belonging to the A-cluster a.
$\int dN(\mu, \Sigma)$	An integral over the density function of the (possibly multivariate) normal distribution with parameters μ and Σ.

Table A.1: Notation summary.

Bibliography

Baumgartner, H., and Steenkamp, J. B. E. M. (2001), "Response Styles in Marketing Research: A Cross-National Investigation," *Journal of Marketing Research*, 38, 143-156.

Celeux, G., and Govaert, G. (1995), "Gaussian Parsimonious Clustering Models," *Pattern Recognition*, 28, 781-793.

Cronbach, L. J. (1946), "Response Set and Test Validity," *Educational and Psychological Measurment*, 6, 475-494.

De Jong, M. G., Steenkamp, J. B. E. M., Fox, J. P., and Baumgartner, H. (2008), "Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation," *Journal of Marketing Research*, 45, 104-115.

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

Dolnicar, S., Grün, B., and Rossiter, J. R. (2010), "An initial empirical guide to translating between different answer formats," in *European Marketing Academy (EMAC) Conference Proceedings*, Copenhagen, Denmark: Copenhagen Business School.

Fraley, C., and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611-631.

Javaras, K. N., and Ripley, B. D. (2007), "An 'Unfolding' Latent Variable Model for Likert Attitude Data: Drawing Inferences Adjusted for Response Style," *Journal of the American Statistical Association*, 102, 454-463.

Johnson, T. R. (2003), "On the Use of Heterogeneous Thresholds Ordinal Regression Models to Account for Individual Differences in Response Style," *Psychometrika*, 68, 563-583.

Johnson, T. R., and Bolt, D. M. (2010), "On the Use of Factor-Analytic Multinomial Logit Item Response Models to Account for Individual Differences in Response Style," *Journal of Educational and Behavioral Statistics*, 35, 92-114.

Jöreskog, K. G., and Moustaki, I. (2001), "Factor Analysis of Ordinal Variables: A Comparison of Three Approaches," *Multivariate Behavioral Research*, 36, 347-387.

Lee, S.-Y., Poon, W.-Y., and Bentler, P. M. (1990), "Full Maximum Likelihood Analysis of Structural Equation Models with Polytomous Variables," *Statistics & Probability Letters*, 9, 91-97.

Lenk, P., Wedel, M., and Böckenholt, U. (2006), "Bayesian Estimation of Circumplex Models Subject to Prior Theory Constraints and Scale-Usage Bias," *Psychometrika*, 71, 33-55.

McLachlan, G. J., and Peel, D. (2000), *Finite Mixture Models*, New York: Willey.

Paulhus, D. L. (1991), "Measurement and Control of Response Bias," in *Measures of Personality and Social Psychological Attitudes*, eds. J. P. Robinson, P. R. Shaver, and L. S. Wrightman, San Diego: Academic Press, pp. 17-59.

R Development Core Team (2009), *R: A language and environment for statistical computing*, Vienna: R Foundation for Statistical Computing, URL: `http://www.r-project.org`.

Rossi, P. E., Gilula, Z., and Allenby, G. M. (2001), "Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach," *Journal of the American Statistical Association*, 96, 20-31.

Van Rosmalen, J., van Herk, H., and Groenen, J. F. (2010), "Identifying Response Styles: A Latent-Class Bilinear Multinomial Logit Model," *Journal of Marketing Research*, 47, 157-172.

Watkins, D., and Cheung, S. (1995), "Culture, Gender, and Response Bias: An Analysis of Responses to the Self-Description Questionnaire," *Journal of Cross-Cultural Psychology*, 26, 490-504.

Wolfe, R., and Firth, W. (2002), "Modelling Subjective Use of an Ordinal Response Scale in a Many Period Crossover Experiment," *Journal of the Royal Statistical Society*, Ser. C, 51, 245-255.