On heterogeneous latent class models with applications to the analysis of rating scores

Aurélie Bertrand* Christian M. Hafner†

June 20, 2012

Abstract

Discovering the preferences and the behaviour of consumers is a key challenge in marketing. Information about such topics can be gathered through surveys in which the respondents must assign a score to a number of items. In this article we suggest a strategy to analyze such data and achieve this objective: it consists in identifying groups of consumers whose response patterns are similar and characterizing them in terms of preferences and covariates. We use latent class models allowing for heterogeneity of both latent class and within-class probabilities across individuals. We illustrate the proposed methodology using data about the preferences of Belgian households for supermarkets.

Keywords: Latent class analysis, rating scores, heterogeneity, EM algorithm, marketing

JEL classification: C35, C38, C87, M31

*ISBA, Université catholique de Louvain

†Corresponding author. ISBA and CORE, Université catholique de Louvain, Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium; christian.hafner@uclouvain.be

1 Introduction

According to the Direct Marketing Association[1], direct marketing is *an interactive process of addressable communication that uses one or more advertising media to effect, at any location, a measurable sale, lead, retail purchase or charitable donation, with this activity analyzed on a database for the development of ongoing mutually beneficial relationships between marketers and customers, prospects or donors.* Such an approach leads to more effective campaigns, by targeting individuals most likely to be interested in a specific offer. However, this strategy requires much more information about all potential customers, since one has to understand their current behaviour and preferences in order to be able to predict them in the future. Hence the use of databases gathering information about individuals and on which data mining algorithms can be trained. This data can relate, among others, to the profile of the individuals and their preferences (e.g. for a brand, for specific product attributes). This information can be collected by conducting a survey, which can be administered in different ways, as explained in Malhotra et al. (2007): by phone, by personal discussion (in the home of the respondent, in a mall, in the street, etc.), by mail, by email and on the internet.

In this paper, we consider the generic context of a self-administered study containing two main types of questions:

Manifest variables Rating questions, i.e. questions in which respondents must assign a score, on a given ordinal scale, to each item of interest. In our context, such questions can typically refer to opinions or behaviours (*I am willing to pay more for environment-friendly products: Strongly agree - Agree - Neither agree nor disagree - Disagree - Strongly disagree*) or frequency of behaviours (*How often do you shop in supermarket x: Never - From time to time - Regularly - Most often*).

Concomitant variables Two main types of individual characteristics: the profile of the respondent (age, gender, profession, city, income, etc.) and additional variables, which can be discrete (nominal or ordinal) or continuous and can concern many different characteristics, behaviours or preferences of the respondent.

This paper introduces a methodology, based on latent class analysis (LCA), that allows us to identify groups of individuals (characterized by some concomitant variables) with a similar pattern of responses to the manifest variables, groups of items which tend to be chosen together, and links between the individual profiles and the response patterns. In comparison with some common data mining algorithms, this methodology emphasizes the understanding of the obtained clusters of individuals, rather than the prediction of the cluster membership. We introduce heterogeneity across individuals and allow latent class probabilities to depend on covariates, which delivers detailed characterizations of the clusters.

Since the most complete model that we consider allows for additional within-class heterogeneity, it enables us to take into account, to a certain extent, an issue related to rating scales: the potential existence of response styles. According to van Herk et al. (2004), citing Paulhus (1971), *response bias is a systematic tendency to respond to a range of questionnaire items on some other basis than the specific item content*, and *if a respondent displays bias consistently across items and methods, this bias is called a response style.* In the context of

[1] http://www.the-dma.org

psychological cross-cultural studies, the authors make a distinction between three types of response bias: socially desirable responding (influence of the cultural norms), acquiescence (*the tendency to agree rather than disagree with items*) and extreme response bias (*the tendency to endorse extreme response categories on a rating scale (e.g., the 1 and/or 5 on a 5-point scale), regardless of content*). When such response styles are present, if they are not taken into account, the analysis of the ratings can yield misleading results. In this work, we could encounter the last two types, since we do not consider cross-cultural studies. We can indeed expect some respondents to always use the same part of the rating scale (leading to differences in the score mean across respondents). Moreover, there can also be differences in the dispersion of the ratings given by the respondents (there are thus differences in the score variance across individuals). If this response bias can be explained by some known covariates, then the model allowing for within-class heterogeneity addresses this issue.

The next section describes the methodology that we suggest in order to analyze rating questions. We first briefly introduce the standard latent class model described, among others, by Goodman (1974). We then present two of its extensions which are useful in our context: the latent class model with covariates, which is an extension of the model proposed by Bandeen-Roche et al. (1997) to the case of polytomous variables, and the latent class model with covariate effects on underlying and measured variables of Huang and Bandeen-Roche (2004). We provide a function in R that deals with estimation and inference of this model. We then propose a strategy for model specification, including the choice of the number of classes, selection of covariates, assessment of identifiability and diagnostic tests.

In Section 3, we apply this methodology to a marketing database about the preferences of Belgian households for supermarkets. We are interested in ordinal variables representing the shopping frequency reported by the household on the scale *Never - From time to time - Regularly - Most often* for 7 supermarkets.

Finally, in Section 4, we expose the most important marketing and statistical conclusions that can be drawn from our results. We also discuss the limitations of our approach and suggest some ideas for future work.

2 Heterogeneous latent class models

In latent class analysis, we would like to use several categorical indicators to measure an unobservable concept which is conceptually defined but cannot be measured directly (see e.g. Bandeen-Roche et al., 1997). A description of the standard latent class analysis model can be found, for example, in McCutcheon (1987) and is summarized here. This method, which can be seen as a categorical data analogue to factor analysis, allows one to analyze the structure of the relationships among *manifest*, i.e. observed, discrete (nominal or ordinal) variables, in order to characterize a categorical *latent*, i.e. unobserved, variable. More precisely, it allows one to identify exclusive latent classes which explain the distribution of the observations in the contingency table obtained by crossing the manifest variables. This corresponds to our objective: identify groups of individuals that explain the observed response patterns.

The basic assumption of latent class analysis is that the observed variables are caused by a latent categorical variable: they are (imperfect) indicators of this unobservable variable. Relationships between these manifest variables can be either causal (an independent variable causes a dependent variable) or symmetrical. It is also assumed that the covariation existing

between the observed variables is explained by the latent variable: the relation exists because the manifest variables are all linked to the latent one. Hence, within each class of the latter variable, the covariation between manifest variables should not be higher than random covariation: this is the *conditional independence* (or *local independence*) condition. The goal is then to find classes within which the observed variables are independent. Neither the assumption of multivariate normality nor that of continuity of measurement are required.

2.1 The standard latent class model

If we denote

- individuals by i, where $i = 1, \ldots, N$,
- manifest variables (items) by Y_m, where $m = 1, \ldots, M$,
- levels of the manifest variables by k, where $k = 1, \ldots, K_m$ and $m = 1, \ldots, M$,
- the latent variable by S, $S = j$ and $j = 1, \ldots, J$,

then the standard latent class model has the following form:

$$P(\mathbf{Y_i} = \mathbf{y}) = P(Y_{i1} = y_1, \ldots, Y_{iM} = y_M) = \sum_{j=1}^{J} \pi_j \prod_{m=1}^{M} \prod_{k=1}^{K_m} p_{mkj}^{y_{imk}} \tag{1}$$

where $y_{imk} = \mathbf{1}(y_{im} = k)$. The parameters of the model are $\pi_j = P(S_i = j)$ and $p_{mkj} = P(Y_{im} = k|S_i = j)$: the former are the *latent class probabilities* (which describe the distribution of the latent variable) while the latter are the *conditional probabilities* (which allow us to characterize the classes which have been obtained by the method, since they determine the extent to which individuals in a given class are likely to be in each category of the manifest variables).

It is assumed that the model is locally identifiable, as defined and discussed by Bandeen-Roche et al. (1997) and Goodman (1974).

The maximum likelihood estimates of these parameters are found by using the Expectation-Maximization (EM) algorithm, see e.g. McCutcheon (1987). This algorithm is implemented in the R packages `poLCA` (Linzer and Lewis, 2011a) and `randomLCA` (Beath, 2011).

For details on the goodness-of-fit measures, see (among others) Collins et al. (1993) and Formann (2003).

The standard latent class model first allows us to identify (latent) clusters of individuals. Second, for each of these clusters, conditional probabilities of being at each level of each item (given that the individual belongs to a particular latent class) are computed: they enable us to characterize the groups by their general pattern of answers to the manifest variables. However, this model does not characterize the classes in terms of other variables, i.e. it does not relate the latent class probabilities to covariates. This is why, in the strategy that we suggest, we only use this model in order to determine the number of latent classes (as is done in Bandeen-Roche et al., 1997), on the basis of the BIC criterion. The most valuable insights are gained by using two extensions of this standard model: the *latent class regression model* described in Bandeen-Roche et al. (1997) and which we call *latent class model with covariates* in this work (as in Linzer and Lewis, 2011b), and the *latent class model with covariate effects on underlying and measured variables* of Huang and Bandeen-Roche (2004).

2.2 Latent class model with covariates

This model consists in making the latent class membership depend on a set of individual covariates through a linear generalized (multinomial) logit link function. This allows us to summarize the effect of covariates on the underlying mechanism (the latent variable). Bolck et al. (2004) show that when regressing the latent class membership obtained by a standard LCA on the covariates of interest through a multinomial logistic regression, the estimated parameters of the model relating latent class probabilities to covariates can be asymptotically biased.

Bandeen-Roche et al. (1997) originally consider binary manifest variables, but we present the extension to polytomous variables, assuming K_m categories for manifest variable m.

The model is the following ($\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})$ is the $(P+1) \times 1$ individual vector of covariates, which can be categorical or continuous):

$$P(\mathbf{Y_i} = \mathbf{y}|\mathbf{x}_i) = P(Y_{i1} = y_1, \ldots, Y_{iM} = y_M|\mathbf{x}_i) = \sum_{j=1}^{J} \pi_j(\mathbf{x}_i'\boldsymbol{\beta}) \prod_{m=1}^{M} \prod_{k=1}^{K_m} p_{mkj}^{y_{imk}}, \quad i = 1, \ldots, N$$

$$\begin{aligned} \log\left(\frac{\pi_j(\mathbf{x}_i'\boldsymbol{\beta})}{\pi_J(\mathbf{x}_i'\boldsymbol{\beta})}\right) &= \beta_{0j} + \beta_{1j}x_{i1} + \ldots + \beta_{Pj}x_{iP} \\ &= \mathbf{x}_i'\boldsymbol{\beta}_j \qquad i = 1, \ldots, N; \quad j = 1, \ldots, (J-1) \end{aligned}$$

Two conditions are linked with this model, as pointed out by Bandeen-Roche et al. (1999): no response associations within individuals (*conditional independence*) and no direct effects of covariates on the manifest variables, i.e. the conditional probabilities do not depend on covariates (*nondifferential measurement*): $P(Y_{im} = y_m|S_i, \mathbf{x}_i) = P(Y_{im} = y_m|S_i), m = 1, \ldots, M$. This implies that $\mathbf{Y}_i$ and $\mathbf{x}_i$ are independent given S_i.

The authors advise to fit standard latent class models without covariates, then choose the "optimal" number of classes and finally fit latent class models with covariates with this number of classes.

Bandeen-Roche et al. (1997) use the EM algorithm to estimate all parameters: this approach is implemented in the R package `poLCA` (Linzer and Lewis, 2011a), which handles polytomous manifest variables. Details about the local identifiability issue and the goodness of fit of the model can be found in Bandeen-Roche et al. (1997).

2.3 Latent class model with covariate effects on underlying and measured variables

In the generic case that we are studying (individuals giving ratings to a number of items), it is possible that we will observe response styles (as mentioned in the introduction). If the behaviour can be explained by the covariates used to predict class membership, the nondifferential measurement assumption does not hold.

In order to analyze self-reported scorings of items, Huang and Bandeen-Roche (2004) use a latent class model with covariate effects on latent and manifest variables, i.e. they impose logit-type models (through the use of individual-level covariates) on both the latent class and the conditional probabilities. They argue that letting the within-class distribution of manifest variables vary among individuals aims to adjust for characteristics (other than the latent classes) that determine manifest variables, hence trying to improve the accuracy of

the classification of individuals. This model thus lets the probability of having assigned a specific rating to a specific item, given the latent class, depend on known covariates. If some of these covariates cause respondents to assign specific ratings more frequently (thus leading to differences in the mean and/or the variance of ratings across respondents), then this model will correct for it.

The authors make the distinction between two sets of covariates:

- $\mathbf{x_i} = (1, x_{i1}, \ldots, x_{iP})^T$ are assumed to be related with the latent class probabilities.
- $\mathbf{z_i} = (\mathbf{z_{i1}}, \ldots, \mathbf{z_{iM}})$ with $\mathbf{z_{im}} = (1, z_{im1}, \ldots, z_{imL})^T$, $m = 1, \ldots, M$ are covariates which can have a direct effect on the manifest variables. These covariates can be specific to each manifest variable (hence the subscript m).

These covariates can be continuous and/or discrete, and the two sets can be mutually exclusive or overlap.

The regression extension of latent class analysis is then (the authors use the logit link function):

$$\begin{aligned}
P(\mathbf{Y_i} = \mathbf{y}|\mathbf{x}_i, \mathbf{z}_i) &= P(Y_{i1} = y_1, \ldots, Y_{iM} = y_M|\mathbf{x}_i, \mathbf{z}_i) \\
&= \sum_{j=1}^{J} \left\{ \pi_j(\mathbf{x}_i'\boldsymbol{\beta}) \prod_{m=1}^{M} p_{mkj}^{y_{imk}} (\boldsymbol{\gamma}_{mj} + \boldsymbol{z}_{im}'\boldsymbol{\alpha}_m) \right\} \\
\log\left(\frac{\pi_j(\mathbf{x}_i'\boldsymbol{\beta})}{\pi_J(\mathbf{x}_i'\boldsymbol{\beta})}\right) &= \beta_{0j} + \beta_{1j}x_{i1} + \ldots + \beta_{Pj}x_{iP} \\
&= \mathbf{x}_i'\boldsymbol{\beta}_j \qquad i = 1, \ldots, N; \quad j = 1, \ldots, (J-1) \qquad (2) \\
\log\left(\frac{p_{mkj}(\boldsymbol{\gamma}_{mj} + \mathbf{z}_{im}'\boldsymbol{\alpha}_m)}{p_{mKj}(\boldsymbol{\gamma}_{mj} + \boldsymbol{z}_{im}'\boldsymbol{\alpha}_m)}\right) &= \gamma_{mkj} + \alpha_{1mk}z_{im1} + \ldots + \alpha_{Lmk}z_{imL} \\
&= \gamma_{mkj} + \mathbf{z}_{im}'\boldsymbol{\alpha}_{mk} \qquad (3)
\end{aligned}$$

$$i = 1, \ldots, N; \quad m = 1, \ldots, M; \quad k = 1, \ldots, (K_m - 1); \quad j = 1, \ldots, J$$

If all regression coefficients α and β are set to 0, this model reduces to basic LCA, while setting $\gamma_{mkj} = \pm\infty$ corresponds to a constrained model with the corresponding conditional probability being 0 or 1.

The coefficients α_{lmk} are not allowed to vary across latent classes because, otherwise, model identifiability could fail. Theorem 1 of Huang and Bandeen-Roche (2004) gives sufficient conditions for local identifiability.

Huang and Bandeen-Roche (2004) present the three assumptions underlying this model:

1. Latent class probabilities are associated with $\mathbf{x}_i$ only,

$$P(S_i = j|\mathbf{x}_i, \mathbf{z}_i) = P(S_i = j|\mathbf{x}_i)$$

2. Responses conditional on class membership only depend on $\mathbf{z}_i$,

$$P(Y_{i1} = y_1, \ldots, Y_{iM} = y_M|S_i, \mathbf{x}_i, \mathbf{z}_i) = P(Y_{i1} = y_1, \ldots, Y_{iM} = y_M|S_i, \mathbf{z}_i)$$

3. Manifest variables are independent given class membership and $\mathbf{z}_i$,

$$P(Y_{i1} = y_1, \ldots, Y_{iM} = y_M | S_i, \mathbf{z}_i) = \prod_{m=1}^{M} P(Y_{im} = y_m | S_i, \mathbf{z}_{im})$$

In the context of latent class analysis, another class of models can adjust for differences in the mean ratings across respondents: *latent class models with random effects*, described in Qu et al. (1996). They take into account conditional dependence, which arises when similarity among responses to several items is caused by some individual-specific characteristics besides the latent class. The idea is that the individual characteristics can be summarized by a continuous random variable T, which is unobservable. That random variable varies from person to person and is normally distributed with zero mean and unit variance. The conditional probability of a certain rating to an item, given the latent class and the random effect, is assumed to depend on S_i and T through a regression model. This model can be estimated with the R package `randomLCA` (Beath, 2011). Compared with the latent class model with covariate effects on underlying and measured variables, this model with random effect does not require covariates to model the response bias. On the other hand, it does not allow to include covariates to explain the latent class membership.

2.4 Implementation in R of the latent class model with covariate effects on underlying and measured variables

As no R package currently deals with models with covariate effects on both the conditional and the latent class probabilities, we provide it for download on our webpage[2], which is an extension of the source code of the R package `poLCA` (Linzer and Lewis, 2011a) to the methodology proposed by Huang and Bandeen-Roche (2004).

The main function is `covLCA()`. Its output contains, among others, the estimated parameters ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) as well as their covariance matrix, the estimated conditional probabilities evaluated at the sample mean of the covariates, and the respective size of each latent class. In the current version, all manifest variables must have the same number of categories, i.e. $K_m = K \quad \forall m$ and the function retains only individuals with no missing values in both manifest variables and covariates.

In order to facilitate convergence, the function first computes initial estimates for the parameters, by applying the idea from Huang and Bandeen-Roche (2004). For the parameters $\boldsymbol{\beta}$, a standard latent class model is estimated, then each individual is assigned to a class with the posterior probabilities of class membership from this model, and finally a multinomial logistic regression model relating the latent classes to the covariates x is fitted, whose coefficient estimates give initial estimates of $\boldsymbol{\beta}$. As far as the two other sets of parameters are concerned, M different multinomial logistic regression models are fitted for $(Y_{i1}, \mathbf{z}_{i1})$, ..., $(Y_{iM}, \mathbf{z}_{iM})$ and the corresponding estimated coefficients are taken as initial values for parameters α and γ.

The E and M-step are then iterated (the M-step being implemented for each subset of parameters) until the improvement in the log-likelihood becomes smaller than a chosen threshold (10^{-10} by default). The function is provided with the analytical expressions of all derivatives.

[2] `http://www.uclouvain.be/christian.hafner`

2.5 A strategy for the analysis of rating scores

On the basis of the papers of Bandeen-Roche et al. (1997 and 1999) and Huang and Bandeen-Roche (2004), we suggest the following strategy for the analysis of rating scores. We work in the R environment.

2.5.1 Choice of the number of latent classes

We begin by estimating some standard latent class models without covariates, each of them with a different number of latent classes. We use the R function `poLCA()` (Linzer and Lewis, 2011a) and estimate each model 20 times with different initial values, in order to try to find the global maximum (and not a local one). We then compare the obtained models on the basis of the AIC and BIC criteria (computed by R). We decide to choose the model with the lowest BIC value. We can also assess the performance of each model using statistics such as the likelihood ratio chi-squared statistic and the Pearson goodness-of-fit statistic, provided that the data matrix is not sparse. Finally, we choose one single model, which we denote by LCA1.n, whose number of latent classes will be used in the subsequent analyses.

We must assess the local identifiability of this model: we implement Goodman's (1974) sufficient condition. First, the number of estimated parameters can not exceed the number of independent observations. Second, we have to check whether the rank of the matrix containing the derivatives of the joint probability $P(Y_1 = y_1, \ldots, Y_M = y_M) = \sum_{j=1}^{J} P(Y_1 = y_1, \ldots, Y_M = y_M, S = j)$ with respect to all parameters has full column rank. If we denote by h a manifest variable pattern, $h = 1, \ldots, (\prod_{m=1}^{M} K_m) - 1$, rows of the Jacobian correspond to profiles and columns correspond to parameters. The profile probabilities are:

$$\begin{aligned} P_h &= P(Y_1 = y_{h1}, \ldots, Y_M = y_{hM}) \\ &= \sum_{j=1}^{J} \pi_j \prod_{m=1}^{M} \left\{ \prod_{k=1}^{K_m - 1} p_{mkj}^{y_{hmk}} \left(1 - \sum_{k=1}^{K_m - 1} p_{mkj} \right)^{y_{hmK_m}} \right\}, \quad h = 1, \ldots, \prod_{m=1}^{M} K_m - 1 \end{aligned}$$

where y_{hm} represents the value taken by manifest variable m in profile h and $y_{hmk} = \mathbf{1}(y_{hm} = k)$. The derivatives with respect to latent class probabilities are:

$$\frac{\partial P_h}{\partial \pi_j} = \prod_{m=1}^{M} \prod_{k=1}^{K_m} p_{mkj}^{y_{hmk}} - \prod_{m=1}^{M} \prod_{k=1}^{K_m} p_{mkJ}^{y_{hmk}}, \quad h = 1, \ldots, \prod_{m=1}^{M} K_m - 1; \quad j = 1, \ldots, J-1$$

while the derivatives with respect to conditional probabilities are:

$$\frac{\partial P_h}{\partial p_{nlj}} = \begin{cases} 0 & \text{if } y_{hnl} = 0 \text{ and } y_{hnK_n} = 0 \\ \pi_j \prod_{m=1; m \neq n}^{M} \prod_{k=1}^{K_m} p_{mkj}^{y_{hmk}} & \text{if } y_{hnl} = 1 \text{ and } y_{hnK_n} = 0 \\ -\pi_j \prod_{m=1; m \neq n}^{M} \prod_{k=1}^{K_m} p_{mkj}^{y_{hmk}} & \text{if } y_{hnl} = 0 \text{ and } y_{hnK_n} = 1 \end{cases}$$

$$h = 1, \ldots, \prod_{m=1}^{M} K_m - 1; \quad n = 1, \ldots, M; \quad l = 1, \ldots, K_n - 1; \quad j = 1, \ldots, J$$

We can compute the rank of this matrix using, for example, the R function `qr()` in package `base` (R Development Core Team, 2010).

As noted by Reboussin et al. (2008), conditional dependence between manifest variables can lead to goodness-of-fit statistics preferring models with more latent classes. If we suspect such a problem, we can first carry out some diagnostics for conditional dependence on model LCA1.n in order to verify whether the assumption of conditional independence holds. If it does, we may want to check whether conditional independence is artificially attained by assuming a too large number of classes, some of them being not very meaningful. We can then estimate some *latent class models with random effects*, described in Qu et al. (1996) and mentioned in Section 2.3. We estimate these models using the R package `randomLCA` (Beath, 2011) and see whether the number of latent classes chosen by both approaches (with and without a random effect) agree. If this is the case, this number of classes can be used in all subsequent models. If both approaches lead to different conclusions, and if the most parsimonious model is easily interpretable while the most complete model seems to contain too many classes, the former should be chosen.

2.5.2 Introduction of covariates to model the latent class probabilities

We then fit a latent class model relating the latent class probabilities to some covariates, with the number of latent classes chosen in the previous step: this model is obtained with `poLCA()` (Linzer and Lewis, 2011a) and is denoted by LCA2.n. In order to choose which covariates to include in this model, we perform some chi-squared Pearson's tests on the table crossing the latent class membership from LCA1.n with the potential covariates. We must take care that the total number of parameters to be estimated can not exceed the number of free cells of the contingency table. In order to simplify convergence of the estimation procedure and the finding of the global maximum, we provide the R function with the estimated conditional probabilities of LCA1.n as starting values for the estimates of the conditional probabilities of LCA2.n. We check whether the introduction of covariates significantly improves the adjustment quality (i.e. whether there is an association between the covariates and the latent variable), on the basis of a likelihood ratio test comparing the two nested models LCA1.n and LCA2.n. If we include more than one covariate, we also test the overall relationship between the latent class membership and each covariate through another likelihood ratio test, before looking at the effect of this covariate on the distinction between two classes.

To assess the overall quality of the fit of LCA2.n, we use the diagnostic suggested in Bandeen-Roche et al. (1997): we first compare the observed proportions in each category of each manifest variable to those predicted by model LCA2.n, for each pattern of covariates. In R, this analysis can be conducted as follows: the computation of observed proportions as a function of a covariate is straightforward, while the predicted proportions for the different levels of covariates are:

$$\hat{P}(Y_m = k|\mathbf{x}) = \sum_{j=1}^{J} \hat{P}(Y_m = k|S = j)\hat{P}(S = j|\mathbf{x})$$

In order to check whether the assumption of nondifferential measurement holds, we apply the algorithm suggested in Bandeen-Roche et al. (1997). We first randomly assign each individual to one latent class based on the posterior latent class probabilities of LCA2.n. Then we create a new categorical variable representing the manifest pattern (i.e. a particular combination of the M manifest variables among the $\prod_m K_m$ possible ones) and regress, in

each latent class, this pattern on the covariates through a multinomial logistic model by using, for example, the R package `mlogit` (Croissant, 2011). Significant associations show that the first assumption is violated.

In the case of *binary* manifest variables, the assumption of conditional independence can be assessed by applying Uebersax's (2000) method which consists in constructing, for each pair of manifest variables, the observed and predicted two-way frequency tables, then computing the log-odds ratio in both of them and finally, calculating a score comparing the expected and observed log-odds ratios:

$$z = \frac{LOR_{obs} - LOR_{pred}}{\sqrt{(1/n_{00} + 1/n_{10} + 1/n_{01} + 1/n_{11})}} \tag{4}$$

where n_{ij} is the observed frequency at the intersection of category i of the first variable and category j of the second variable, $LOR_{obs} = \log(n_{00}n_{11}/n_{01}n_{10})$, and LOR_{pred} is the corresponding predicted log-odds ratio. If $|z|$ is greater than a chosen critical value (1.645 or 1.96 for example), there is evidence that the items are conditionally dependent. The z-values should only be considered as guides, the most important being to compare their relative sizes. This method is used by Reboussin et al. (2008).

As far as the local identifiability of LCA2.n is concerned, we use Theorem 1 in Bandeen-Roche et al. (1997): as the local identifiability of model LCA1.n has already been checked, this theorem only requires the full rank of the design matrix of covariates, which is easily computed in R, and the assumption that the vector of latent class probabilities has no zero element for at least one individual, which can be checked after the model has been estimated.

2.5.3 Introduction of covariates to model the conditional probabilities

We can now fit LCA3.n, a latent class model relating both the conditional and the latent class probabilities to covariates, and having the same number of classes as chosen in the first step.

In order to choose which covariates to link to the conditional probabilities, we can apply the same method as the one used to assess the nondifferential measurement assumption of LCA2.n: in each latent class, we regress the manifest pattern on the potential covariates.

The estimation of LCA3.n is carried out using our extension of package `poLCA` (Linzer and Lewis, 2011a), function `covLCA()`. As for the two previous models, we run it several times in order to find the global maximum of the log-likelihood. We compare the model fit with that of model LCA2.n, using the same technique as in the previous step: we compare proportions in each manifest variable predicted by LCA3.n to those observed and to those predicted by LCA2.n. A likelihood ratio test allows one to determine whether the additional covariate(s) lead(s) to a significant improvement in the model fit. We also verify whether this model is locally identifiable: we apply the theorem introduced in Huang and Bandeen-Roche (2004) (and implemented in our R program).

3 Application to rating scores in marketing

Our application is made in collaboration with the Managed Marketing Services (henceforth MMS) unit of Business & Decision Brussels, a consulting company also working in the fields of Business Intelligence, Customer Intelligence, Information Management, Risk Management,

Human capital Management, e-Business and Life Sciences[3]. The MMS unit works on a direct marketing database, result of a self-administered survey sent by mail. This database contains various information about more than 710 000 Belgian households: who they are, where they live, their car, their house (presence of a swimming pool, number of bedrooms), their consumption patterns, their hobbies, their holidays, etc.

In this work, we analyze the preferences for supermarkets of Belgian households living in one of the Belgian provinces, Walloon Brabant. In this context, the manifest variables are 7 ordinal variables representing the shopping frequency reported by the household on the scale *Never - From time to time - Regularly - Most often*, for each of the following supermarkets:

- Aldi, a German discounter.
- Carrefour and GB: hypermarkets Carrefour and supermarkets GB are two of the store brands of the French supermarket chain Carrefour. Carrefour bought out GB in 2000.
- Colruyt, a Belgian discounter.
- Cora, a Belgian supermarket chain. In 1967, the first supermarket opened under the name "Carrefour". The name "Cora" appeared only in 1974.
- Delhaize, a Belgian supermarket chain which owns several store brands, among them: Delhaize Le Lion, AD Delhaize, Red Market, City, Proxy and Shop'n Go.
- Lidl, a German discounter.

These supermarkets can be grouped into three categories: discounters which sell (hardly) no national brands (Aldi and Lidl), a discounter which sells both national and its own store brands at low prices (Colruyt), and supermarkets selling national and store brands which do not particularly emphasize low prices (Carrefour, GB, Cora and Delhaize).

We would like to be able to identify relationships between household profiles and supermarkets, which would allow us to characterize the most frequent customers of each supermarket, identify groups of supermarkets which are often visited by the same households, and predict the preferred supermarkets of a household for which the available information is limited. From the point of view of latent class analysis, the latent variable can be broadly defined as the preference for some attributes relating to supermarkets (which divide households into groups), with the indicators being the frequency with which one shops in each of these places.

3.1 Choice of the manifest variables and preliminary data treatment

In marketing surveys using rating scales, the respondents can differ in their mean rating (as explained in the introduction), in the sense that some households always assign high scores, while others only use the lower part of the rating scale. Hence, the absolute level of each variable is not very meaningful. If the response styles can be explained by some known covariates, then the latent class model with covariate effects on underlying and measured variables will take them into account. However, we suspect that the covariates that we consider in this work can not fully explain this phenomenon. It could thus be interesting to consider the *relative* scores given by each respondent. This can be achieved by replacing,

[3]http://www.businessdecision.be/

for each individual, the *rating* of each item by its *ranking* (among the items rated by this respondent): the item(s) with the highest rating given by this respondent is (are) assigned the value 1, the item(s) with the second highest rating from this respondent receive(s) the value 2, and so on.

In our specific study, we want to identify clusters of individuals on the basis of their favorite supermarket(s), and not on the basis of their whole preference pattern. We therefore decide to apply a simple preliminary transformation to our data: we create, for each categorical manifest variable, a corresponding dichotomous variable (taking on values 1 and 2) indicating whether this supermarket has been assigned the highest score attributed by the household: if we denote the original ratings by Y^*_{im} and the new binary indicators by Y_{im}, this transformation is of the form $Y_{im} = 1 + \mathbf{1}(Y^*_{im} = \max_m Y^*_{im})$, $i = 1, \ldots, N$, $m = 1, \ldots, M$, hence $k \in \{1, 2\}$ and $K_m = 2, \forall m$. A household which has assigned its highest score to more than one supermarket thus receives the value 2 for all these supermarkets. These new variables can be interpreted as the favourite supermarkets: they represent the relative scores given by each household and allow one to analyze relative preferences. The binarization can be seen as a simple ranking (the preferred item(s) versus the other ones) and hence addresses the issue of the differences in mean of the ratings across individuals.

As far as the covariates are concerned, we decide to consider four of the variables of the database:

- Age of the respondent, $X_1 \in \{< 18, 18 - 35, 35 - 45, 45 - 55, 55 - 65, 65 - 75, > 75\}$
- Gender of the respondent, $X_2 \in \{M, F\}$
- Socioprofessional activity of the respondent, $X_3 \in$ {Retired, Looking for a job, Student, Civil servant, House wife/husband, Employed in the private sector, Self-employed, Other}
- Highest education level in the household, $X_4 \in$ {Secondary school, Graduat, Licence/ maitrise/ doctorat, Other}

These four covariates are coded as nominal variables using dummy variables.

There are missing values in all variables: we retain only the households with no missing values in both the manifest variables and the four covariates, and which have assigned a score higher than "never" at least once: 5473 households among 18893 are retained.

3.2 Choice of the number of latent classes

The BIC criterion favors a standard latent class model with nine latent classes: this model is locally identifiable. Seven classes are easily interpretable, but the two other ones (which represent together less than 10% of the households) group all households whose preferences do not appear very clearly (households which have different behaviours). The diagnostics for conditional dependence, given in equation (4) and applied to the models with eight (first lines in Table 1) and nine latent classes (not displayed here) show that at least nine latent classes are necessary to model the dependence existing in the data: in LCA1.8, there seems to be conditional dependence between Delhaize and Colruyt. The crucial question is whether there are indeed nine "behaviour patterns" in the population, or whether the results are influenced by a potential dependence within households.

Table 1: Diagnosis for conditional independence for LCA1.8 (1st lines), LCA2.8 (2nd lines) and LCA3.8 (3rd and 4th lines, for men and women respectively): scores comparing observed and predicted log-odds ratios in each contingency table relating two manifest variables. The scores higher than 2 are in bold.

	Aldi	Carrefour	Colruyt	Cora	Delhaize	GB	Lidl
Aldi		0.19	0.30	-0.50	-0.22	0.85	0.96
		0.70	1.01	-0.55	0.25	0.43	1.03
		0.06	0.65	-0.19	0.11	0.70	**3.09**
		0.43	-0.16	-0.53	-0.27	-0.05	-0.68
Carrefour			1.64	0.65	1.57	0.60	-0.31
			1.49	0.64	1.33	0.49	-0.13
			0.50	-0.01	0.55	0.18	0.81
			1.59	0.84	1.30	0.63	-0.72
Colruyt				1.13	**2.19**	-0.83	0.31
				1.35	**2.12**	-1.01	-0.39
				-0.08	**2.10**	0.72	1.16
				1.94	1.42	-1.72	-0.55
Cora					0.49	0	1.15
					0.73	0.46	1.36
					-0.71	-0.11	1.04
					1.27	0.51	1.00
Delhaize						0.92	-0.19
						1.28	0.77
						1.28	0.52
						0.01	0.01
GB							-0.04
							0.16
							0.53
							0.26

To answer this question, we estimate latent class models with random effects: on the basis of the BIC criterion, a model with eight latent classes is chosen, i.e. one class less than when using no random effect. This fact seems to confirm that, in a model without random effect, nine classes are necessary to obtain conditional independence because of the presence of a dependence between manifest variables within individuals, which is not explained by the latent membership. Both models assuming eight latent classes (with and without random effect) broadly yield similar results. We thus decide to estimate the following models with eight latent classes.

3.3 Modelling of the latent class probabilities

A Pearson's chi-square test is performed on each of the four contingency tables obtained by crossing the predicted latent class (the class for which the posterior probability $\hat{h}_{ij} =$

$\hat{P}(S_i = j|\mathbf{Y}_i = \mathbf{y}_i)$ is the highest) and one covariate: each covariate is associated[4] with the latent membership (although the association with Gender is weaker). Preliminary analyses highlighted a link between the choice of the supermarket and variables pertaining to the standard of living of the household, like the income, the education level, the profession, the size of the house, the car brand, etc. We therefore choose to make the latent class membership depend on the education level: this variable can be easily observed and considered, to a certain extent, as a proxy for the standard of living.

In comparison with the model LCA1.8, the addition of Education significantly improves the fit of the model (likelihood-ratio statistic: 243.55 with 21 degrees of freedom, the p-value is virtually equal to 0). Moreover, the AIC is reduced from 34661 to 34460 and the BIC decreases from 35077 to 35015.

On the basis of the goodness-of-fit diagnostic, we conclude that the model seems to fit the data very well: for each category of Education, the observed and predicted proportions of having each supermarket among its favourite ones are very close to each other (the largest difference, relative on the estimated probabilities, is 0.01 for an estimated probability of 0.04). However, Table 1 shows that, from the point of view of the conditional independence assumption, the situation is similar to the one in LCA1.8: there could also be conditional dependence between Colruyt and Delhaize. The predicted probabilities needed in the z-scores are computed using $\hat{P}(Y_1 = y_1, Y_2 = y_2) = \sum_j \hat{P}(S_i = j)\hat{P}(Y_1 = y_1|S_i = j)\hat{P}(Y_2 = y_2|S_i = j)$ where $\hat{P}(S_i = j)$ is the average over all individuals in the sample of the posterior latent class probabilities. As far as the nondifferential measurement assumption is concerned, in each class except one[5], there is no significant relationship (based on a likelihood ratio test) between the manifest pattern and Education. Finally, it is verified that LCA2.8 is locally identifiable.

3.4 Modelling of the conditional probabilities

In model LCA2.8, the response pattern significantly depends on the gender of the respondent in two latent classes[6]. The "Gender" variable thus seems to be a good candidate. Moreover, since this variable has only two categories, only a few additional parameters must be estimated (91 instead of 84). However, although this estimated model is locally identifiable and that adding Gender significantly improves the fit of the model (likelihood ratio statistic: 17.74 with 7 degrees of freedom, the p-value is 0.01), BIC increases from 35014.58 to 35057.08. According to the BIC criterion, the more parsimonious model LCA2.8 would be chosen: the increase in the number of parameters to be estimated outweighs the improvement in the log-likelihood. Since the results (conditional probabilities and the regression coefficients of the model for the latent class probabilities) are similar in models LCA2.8 and LCA3.8, we present the main findings of the most flexible model, LCA3.8.

Table 1 contains the diagnostics for conditional dependence, calculated for men and women separately, where the estimated odds ratios are computed using the formula $\hat{P}(Y_1 = y_1, Y_2 = y_2|gender) = \sum_j \hat{P}(S_i = j)\hat{P}(Y_1 = y_1|S_i = j, gender)\hat{P}(Y_2 = y_2|S_i = j, gender)$ where $\hat{P}(S_i = j)$ is the sum over all individuals in the sample of the posterior latent class probabilities. For men, the score for the pair Colruyt-Delhaize is still rather high. Moreover, the score between Aldi and Lidl becomes very high. The situation is better for women: no score

[4] P-value of the Pearson's chi-squared test: $< 2.2 \cdot 10^{-16}$ for Age, Education and Profession, 0.04 for Gender.

[5] In class *GB-Carrefour*: p-value of the likelihood ratio test is < 0.01.

[6] P-value in latent class *Lidl-Aldi*: 0.03 ; p-value in class *Delhaize*: 0.05.

exceeds 2 in absolute value.

As far as the overall goodness of fit of the model is concerned, it can be seen in Table 2 that the observed and predicted item probabilities are close, for every gender-education profile.

Table 3 contains the estimated latent class probabilities (for each latent class, the average, over all individuals of the sample, of the estimated posterior probabilities) and conditional probabilities (calculated for men and women separately). The differences in the estimated conditional probabilities between men and women are limited: we will conclude later, on the basis of Table 6, that the differences for supermarkets Carrefour and Colruyt are significant.

These estimated probabilities allow us to characterize the eight latent classes:

1. *GB-Carrefour*: all households go to GB, *many* of them (more than the overall proportion in the whole sample, last row of table 3) to Carrefour too, and *some* of them (but less than the overall proportion in the whole sample) to Delhaize and Colruyt as well.

2. *Cora*: all households go to Cora, some of them to Delhaize, Colruyt and Carrefour too.

3. *Colruyt*: all households go exclusively to Colruyt.

4. *Delhaize*: all households go (almost exclusively) to Delhaize, some of them to Colruyt as well.

5. *Carrefour*: all households go (almost exclusively) to Carrefour, some of them to Colruyt too.

6. *Aldi-Lidl*: all households to to Aldi, many of them to Lidl too, and some of them to Carrefour, Colruyt and Delhaize as well.

7. *Lidl*: all households go to Lidl, some of them to Colruyt too.

8. *Several*: these households are more likely than the average to go to every supermarket.

There appears to be **seven distinct customer profiles**, while the last latent class contains households whose consuming patterns are not clear and which do not belong to any of the other groups. Although this number of profiles corresponds to the number of supermarkets in the study, there are not only one-to-one relationships between the latent classes and the supermarkets: links between some supermarkets are highlighted by the model: between Lidl and Aldi, Delhaize and Colruyt, Cora and Delhaize, Cora and Colruyt, GB and Carrefour. One explanation concerning the latter is that some former GB supermarkets now belong to Carrefour: it is possible that some people fail to distinguish between them. As a result, we can not conclude that there exists a real link between the preference for each of these two supermarkets.

The smallest class is *Cora*, while the two largest groups are *Delhaize* and *Colruyt*, each representing about 20% of the population. This is consistent with the distribution of the preference for each supermarket in the whole sample: the two most cited ones are Colruyt and Delhaize (last row of Table 3).

It appears that in classes *Lidl* and *Aldi-Lidl*, households shopping most often in these places are also likely to visit other supermarkets with the same frequency: we can describe

Table 2: Goodness-of-fit diagnostics for model LCA3.8: for men and women, in each level of Education, observed (first line) versus predicted (second line) proportions of respondents having cited each supermarket. Columns do not sum to 1 as households can cite several supermarkets as their favourite ones.

Gender	Supermarket	Other degree	Secondaire	Graduat	Lic.,mait.,doct.
Men	Aldi	0.34	0.28	0.20	0.17
		0.34	0.29	0.20	0.14
	Carrefour	0.28	0.36	0.32	0.29
		0.31	0.32	0.31	0.30
	Colruyt	0.46	0.41	0.44	0.43
		0.42	0.42	0.44	0.45
	Cora	0.05	0.06	0.04	0.03
		0.04	0.05	0.04	0.04
	Delhaize	0.25	0.34	0.36	0.46
		0.31	0.34	0.36	0.44
	GB	0.20	0.19	0.16	0.14
		0.19	0.19	0.16	0.15
	Lidl	0.23	0.22	0.17	0.13
		0.28	0.24	0.19	0.11
Women	Aldi	0.34	0.29	0.20	0.13
		0.34	0.29	0.20	0.14
	Carrefour	0.26	0.28	0.28	0.27
		0.27	0.29	0.28	0.27
	Colruyt	0.36	0.38	0.42	0.44
		0.39	0.39	0.41	0.42
	Cora	0.04	0.05	0.03	0.03
		0.04	0.05	0.04	0.04
	Delhaize	0.32	0.32	0.37	0.46
		0.31	0.34	0.36	0.44
	GB	0.19	0.19	0.15	0.14
		0.19	0.19	0.15	0.15
	Lidl	0.30	0.26	0.19	0.11
		0.29	0.25	0.19	0.12

Table 3: Estimated conditional $\hat{P}(Y_m = 2|S = j)$ and latent class probabilities for model LCA3.8 for women and men respectively. The reported latent class probabilities are the average (over all individuals) of the latent class probabilities $\hat{P}(S_i = j|\mathbf{x}_i)$. The last row contains the sample proportion, for each supermarket, of households having mentioned this place among their favourite ones: $\frac{1}{N}\sum_i \mathbf{1}(y_{im} = 2)$.

LC		Aldi	Carrefour	Colruyt	Cora	Delhaize	GB	Lidl	LC pr.
GB-Carr.	F	0.08	0.40	0.19	0.01	0.35	1	0.06	0.11
	M	0.07	0.48	0.23	0.01	0.34	1	0.05	
Cora	F	0	0.21	0.31	1	0.35	0	0.06	0.01
	M	0	0.27	0.35	1	0.34	0	0.05	
Colruyt	F	0.03	0.02	1	0	0	0	0	0.22
	M	0.02	0.03	1	0	0	0	0	
Delh.	F	0.02	0.10	0.19	0	1	0	0.02	0.25
	M	0.02	0.13	0.23	0	1	0	0.02	
Carr.	F	0.04	1	0.10	0.01	0	0	0.04	0.12
	M	0.04	1	0.12	0.01	0	0	0.03	
Al.-Lidl	F	1	0.15	0.25	0.01	0.17	0.09	0.52	0.17
	M	1	0.20	0.29	0.02	0.16	0.10	0.49	
Lidl	F	0	0.11	0.23	0	0.10	0.08	1	0.06
	M	0	0.15	0.28	0	0.10	0.09	1	
Several	F	0.80	0.91	0.68	0.39	0.78	0.73	0.72	0.06
	M	0.79	0.94	0.72	0.42	0.77	0.75	0.70	
Global		0.24	0.29	0.41	0.04	0.37	0.17	0.20	

them as ***multi***. One possible explanation is that households buy most of the things they need at Aldi or Lidl, and go to another place to find national brands for specific products. This behaviour can be contrasted with that of the households in the class *Colruyt*: they can be considered as **loyal customers**, since they do not visit other supermarkets as often as the households in classes *Lidl* and *Aldi-Lidl* do, probably because Colruyt offers both low price products and national brands. The situation is similar in the classes *Delhaize* and *Carrefour*: in the former, respondents liking Delhaize rarely have another preferred place (and when they do, they choose Colruyt), and in the latter, households are most likely to visit Colruyt in addition to Carrefour.

We now focus on the part of the model relating the latent class membership to the highest education level of the household. The exponential transformation of the estimated parameters of the multinomial logistic regression model gives the following odds ratios:

$$\exp(\beta_{jp}) = \frac{P(S = j|education = p)/P(S = Several|education = p)}{P(S = j|education = \text{autre})/P(S = Several|education = \text{autre})}$$

The reference category of Education is "Other degree", while the reference class is *Several*, but it can easily be shown that the quotient of the odds ratios pertaining to two latent classes gives the odds ratio comparing these two classes:

$$\frac{\exp(\beta_{jp})}{\exp(\beta_{qp})} = \frac{P(S = j|education = p)/P(S = q|education = p)}{P(S = j|education = \text{autre})/P(S = q|education = \text{autre})}$$

The odds ratios can be found in Table 4. The education level has an influence on the relative probability of being in latent classes *Colruyt*, *Delhaize*, *Aldi-Lidl* and *Lidl* rather than in latent class *Several*. For example, having a *licence*, *maitrise* or *doctorat* makes a household about 3 times more likely (than those with an "other" degree) to be in latent class *Delhaize* rather than in class *Several*, about 2 times more likely to be in latent class *Colruyt* rather than in latent class *Several* and about 3 times less likely to be in latent classes *Aldi-Lidl* and *Lidl* rather than in latent class *Several*. Coefficients for latent classes *GB-Carrefour*, *Carrefour* and *Cora* are not significantly different from 0: we can not conclude that the education level influences the probability of being in one of these classes rather than in class *Several*. To summarize, on the basis of the highest education level of the household, we can broadly identify three sets of classes: (1) *Delhaize* and *Colruyt*, (2) *Cora*, *GB-Carrefour*, *Carrefour* and *Several*, and (3) *Aldi* and *Lidl-Aldi*.

The estimated probabilities of being in each latent class given the highest education level of the household are displayed in Table 5. The estimated probability of membership in latent classes *Colruyt* and *Delhaize* increases with the level of the education level, while the opposite trend can be observed in latent classes *Aldi-Lidl* and *Lidl* (this is similar to what is concluded from LCA2.8).

Overall, the latent class with the smallest probability of membership is *Cora*, while the preferred ones are *Colruyt*, *Delhaize* and *Aldi-Lidl*.

Finally, we consider the model linking the conditional probabilities $P(Y_{im} = 1|S_i = j) = 1 - P(Y_{im} = 2|S_i = j)$ to the gender of the respondent. The estimated coefficients α can be found in Table 6. The only significant differences between men and women exist for the manifest variables Carrefour and Colruyt: the men are more likely than the women to cite these supermarkets. This model should hence improve the accuracy of the classification of the

Table 4: Latent class membership regression parameter estimates from model LCA3.8. The reference category of Education is "Other". "Int." is the intercept, "Sec." is category Secondaire, "Grad." is category Graduat and "Lic." is category Licence,Maitrise,Doctorat.

LC		Coefficient	exp(Coeff.)	Std. error	t value	$P(>\|t\|)$
GB-Carr	(Int.)	0.49	1.63	0.28	1.75	0.088
vs	Sec.	0.06	1.06	0.29	0.20	0.841
Several	Grad.	0.20	1.23	0.31	0.65	0.521
	Lic.	0.29	1.33	0.33	0.86	0.393
Cora	(Int.)	-1.92	0.15	0.54	-3.55	0.001
vs	Sec.	0.40	1.49	0.58	0.68	0.499
Several	Grad.	0.47	1.60	0.61	0.77	0.443
	Lic.	0.82	2.27	0.61	1.35	0.185
Colruyt	(Int.)	1.11	3.04	0.25	4.49	0
vs	Sec.	0.01	1.01	0.26	0.03	0.977
Several	Grad.	0.47	1.59	0.28	1.69	0.100
	Lic.	**0.59**	**1.81**	0.29	2.04	0.049
Delhaize	(Int.)	0.93	2.53	0.25	3.75	0.001
vs	Sec.	0.19	1.21	0.26	0.73	0.468
Several	Grad.	**0.72**	**2.05**	0.28	2.56	0.015
	Lic.	**1.08**	**2.95**	0.29	3.69	0.001
Carr.	(Int.)	0.44	1.56	0.28	1.58	0.123
vs	Sec.	0.12	1.12	0.28	0.41	0.685
Several	Grad.	0.60	1.83	0.30	2.01	0.052
	Lic.	0.56	1.75	0.32	1.77	0.085
Al.-Lidl	(Int.)	1.41	4.11	0.25	5.76	0
vs	Sec.	-0.25	0.78	0.26	-0.94	0.351
Several	Grad.	-0.38	0.68	0.29	-1.33	0.191
	Lic.	**-0.94**	**0.39**	0.32	-2.95	0.005
Lidl	(Int.)	0.27	1.31	0.29	0.95	0.346
vs	Sec.	-0.15	0.86	0.30	-0.51	0.614
Several	Grad.	0.05	1.05	0.32	0.16	0.873
	Lic.	**-0.97**	**0.38**	0.40	-2.41	0.021

Table 5: For each level of Education, estimated probabilities of being in each of the 8 latent classes in model LCA3.8: $\hat{P}(S = j|education = p)$. The modal classes are in bold.

Education	*GB-Carr.*	*Cora*	*Colr.*	*Delh.*	*Carr.*	*Al.-Lidl*	*Lidl*	*Several*	
Autre	0.11	0.01	0.20	0.17	0.10	**0.27**	0.09	0.07	1
Secondaire	0.11	0.01	0.20	0.20	0.12	**0.21**	0.07	0.07	1
Graduat	0.10	0.01	0.24	**0.26**	0.14	0.14	0.07	0.05	1
Lic.,mait.,doct.	0.10	0.02	0.26	**0.35**	0.13	0.08	0.02	0.05	1

Table 6: Estimated effects of being a man (coefficients α) on the conditional probabilities in model LCA3.8

Supermarket	Coefficient	exp(Coeff.)	Std. error	t value	$P(>\|t\|)$
Aldi	0.08	1.09	0.24	0.35	0.732
Carrefour	**-0.34**	**0.71**	0.11	-3.19	0.003
Colruyt	**-0.22**	**0.80**	0.09	-2.56	0.015
Cora	-0.14	0.87	0.24	-0.60	0.554
Delhaize	0.05	1.05	0.13	0.36	0.725
GB	-0.08	0.92	0.21	-0.38	0.703
Lidl	0.11	1.11	0.13	0.78	0.440

households into the latent classes: we expect that some men who were classified in latent class *Colruyt* (resp. *Carrefour*) in LCA2.8 do not belong to this latent class in LCA3.8. Among the 35 respondents belonging to *Colruyt* in LCA2.8 and to *Aldi-Lidl* in LCA3.8, there are 15 men and 20 women: the proportion of men is 43%, higher than the overall proportion of men in the whole sample, 29%. Similarly, among the 40 respondents belonging to *Carrefour* in LCA2.8 but not in LCA3.8, 14 (35%) are men.

4 Conclusions

4.1 Marketing conclusions

The strategy that we suggest allows one to analyze (self-reported or not) ratings of items. More precisely, it identifies groups of individuals with the same pattern of responses, groups of items which often receive the same rating, and links between these two sets of groups. This methodology also takes into account the potential problem of response styles. Thanks to the use of latent class analysis, which yields a large number of parameters, the identified clusters of individuals can be deeply understood. This strategy can thus be useful in many different situations.

In order to illustrate its use, we have applied it to a marketing database about the preferences of Belgian households for supermarkets. We have identified eight groups of customers, seven among which are clearly defined. A distinction has appeared between *loyal* households, i.e. households which shop in only one supermarket with their highest frequency, and *multi* households, i.e. those regularly visiting several different places. The multinomial logistic regression model relating the latent class probabilities to the highest education level of the household has highlighted a significant link between the classes, and hence the favourite supermarket(s), and the covariate. This strong association makes it possible to do some prediction about the favourite supermarket, using only one variable which is easily observed: this can be valuable information for marketing departments.

All these findings are consistent with preliminary analyses.

4.2 Latent class models for the analysis of preferences

One valuable advantage of latent class models in the analysis of ratings is the identification of links between manifest variables: in this context, this feature is of paramount importance,

since the latent classes can contain individuals which do have several favourite items. The simultaneous analysis of each manifest variable independently would not have provided the same information.

Being able to relate the class membership to covariates is a clear improvement of the extensions of the standard latent class model that we have considered. In the application of these models to our database, this highlights a significant link between the highest education level of the household and the classes, hence the favourite supermarket.

However, a limitation of this model must also be mentioned: as latent class analysis involves estimating a large number of parameters, the number of covariates which can be included in the model is limited. Moreover, in our case, with a rather large number of latent classes (Qu et al. (1996) consider 2 classes, Bandeen-Roche et al. (1997), 3 classes, Bandeen-Roche et al. (1999), 5 classes and Huang and Bandeen-Roche (2004), 4 classes), and discrete covariates which have to be replaced by several binary indicators, this restriction is really a handicap. As we wanted to be able to estimate an even more complex model (requiring more parameters), we restricted ourselves to one covariate. As a consequence, although the class membership is, to a certain extent, better understood, such a model does not allow us to make effective prediction. This is the price to pay in order to be able to deeply understand the structure of the latent classes: the large number of conditional probabilities allows one to precisely characterize each group. This is a strength of our approach, which distinguishes it from data mining algorithms, which can be efficient in prediction, but are often used as black boxes.

We can nevertheless imagine a possible way to overcome the complexity problem: we could take the posterior latent class membership provided by a latent class model, and train a data mining algorithm on the same data to predict this nominal variable. We mentioned in Section 2.2 that such an approach leads to biased estimates (compared to the estimation of a model which directly integrates covariates) but, as the introduction of the covariate Education did not deeply modify the interpretation of the latent classes, this approach would be worth trying.

Finally, the most flexible model (the one with covariate effects on both the conditional and the latent class probabilities) seems to offer only marginal improvements in our case, perhaps because the differences in response styles can not be explained very well by the covariates we decided to consider. A latent class model with random effect could be more useful in this situation. On the other hand, we are not aware of the existence of such a model which also integrates covariates to explain the latent class membership. A three-step procedure is also possible, but with the risk of obtaining biased estimates.

Another remark concerns the computational aspects of this study. In the papers presenting the models we have estimated, the sample size is always relatively small (1643 in Bandeen-Roche et al. (1999) for example). In our case, the estimation of the standard latent class models was rapid for a small number of latent classes, and took approximately two hours for LCA1.8 (for 20 runs, with different initial values). The estimation of models with covariate effects on the latent class probabilities was quicker, since it was provided with the results of the other model as initial values. However, estimating model LCA3.8 took 13 hours, which is not really suited to exploratory analyses. Implementing parts of our R function in C would reduce computing time substantially.

4.3 Limitations of our approach and outlook

The methodology we have decided to pursue in order to analyze ratings is one approach among many possible ones. It has several advantages, but some drawbacks too.

The first point which can be brought into question is the family of models we have used, i.e. latent class analysis. We chose it after having reviewed several other methods, but many of them would be worth applying to the type of data we considered. Moreover, latent class models are well suited when the main objective is the understanding of relations between manifest variables. However, as already mentioned, these models are not very efficient for prediction.

Discarding all individuals with missing values in either the manifest variables or the covariates is also questionable since this amounts to losing information and possibly biasing the results (as mentioned by Huang and Bandeen-Roche (2004) in the discussion at the end of their paper). We have namely discovered that people with missing values differ from those without missing value in their age profile[7] (among others).

One possible alternative would consist in replacing the missing values in categorical variables either by one of the existing categories (through a decision tree for example) or by a new category "missing". This would allow us to retain all individuals in the analysis, but this approach could also add noise to the results. As far as the manifest variables (the ratings) are concerned, since we transform them to binary indicators, we could imagine assigning the lowest value to items whose rating is missing, and possibly retaining individuals with no more than a given number of missing values in the original ratings.

Another possibility to deal with missing values suggested by the editor would be to determine, for each individual, the log-likelihood based only on the observed data (and to drop the part depending on missing values).

We also decided to transform the manifest variables into binary indicators, in order to obtain relative (and hence comparable) preferences among households. We could have used the original ordinal ratings (to preserve all information) in a latent class model with random effects (to take into account the dependence between scores assigned by one respondent). However, in such a case, it would be more interesting to treat the ratings as ordinal variables, since they provide more information than nominal ones. The model suggested by Albert (2007), a latent class model with random effects to take into account dependence between *ordinal* manifest variables, could be a solution.

One can also wonder whether the number of latent classes we decided to consider is optimal. On the one hand, the BIC criterion opted for a 9-class model. On the other hand, the BIC value becomes rather stable from 6 latent classes: using an even more parsimonious criterion might result in a model with fewer latent classes.

Finally, in the most flexible models, we have chosen one (and only one) covariate to include on the basis of an independence test between the latent classes from the first model and the different covariates that we were considering. It would be worth finding a stepwise approach able to automatically select the most relevant features from a large set of possible ones (as do, for example, backward and forward selections in multiple linear regression).

[7] P-value of the Pearson's chi-squared test: $< 2.2 \cdot 10^{-16}$.

Acknowledgements

We would like to thank Eric Lecoutre for helpful discussions, and Business & Decision, Brussels, for providing the data.

References

[1] ALBERT P.S. (2007) Random Effects Modeling Approaches for Estimating ROC Curves from Repeated Ordinal Tests without a Gold Standard. *Biometrics*, Vol. 63, 593-602.

[2] BANDEEN-ROCHE K., HUANG G.-H., MUNOZ B., RUBIN G.S. (1999) Determination of Risk Factor Associations with Questionnaire Outcomes: A Methods Case Study. *American Journal of Epidemiology*, Vol. 150, No. 11, 1165-1178.

[3] BANDEEN-ROCHE K., MIGLIORETTI D.L., ZEGER S.L., RATHOUZ P.J. (1997) Latent Variable Regression for Multiple Discrete Outcomes. *Journal of the American Statistical Association*, Vol. 92, No. 440, 1375-1386.

[4] BEATH K. (2011). randomLCA: Random Effects Latent Class Analysis. R package version 0.8-3. http://CRAN.R-project.org/package=randomLCA

[5] BOLCK A., CROON M., HAGENAARS J. (2004) Estimating Latent Structure Models with Categorical Variables: One-Step Versus Three-Step Estimators. *Political Analysis*, Vol. 12, No. 1, 3-27.

[6] COLLINS L.M., FIDLER P.L., WUGALTER S.E., LONG J.D. (1993) Goodness-of-Fit Testing for Latent Class Models. *Multivariate Behavioral Research*, Vol. 28, No. 3, 375-389.

[7] CROISSANT Y. (2011) mlogit: multinomial logit model. R package version 0.2-1. http://CRAN.R-project.org/package=mlogit

[8] FORMANN A.K. (2003) Latent Class Model Diagnosis from a Frequentist Point of View. *Biometrics*, Vol. 59, No. 1, 189-196.

[9] GOODMAN L.A. (1974) Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. *Biometrika*, Vol. 61, No. 2, 215-231.

[10] HUANG G.-H., BANDEEN-ROCHE K. (2004) Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, Vol. 69, No. 1, 5-32.

[11] LANGE K. (1995) A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 57, No. 2, 425-437.

[12] LIN T.H., DAYTON C.M. (1997) Model Selection Information Criteria for Non-Nested Latent Class Models. *Journal of Educational and Behavioral Statistics*, Vol. 22, No. 3, 249-264.

[13] LINZER D. A., LEWIS J. (2011a) poLCA: Polytomous Variable Latent Class Analysis. R package version 1.3.1. http://userwww.service.emory.edu/~dlinzer/poLCA.

[14] LINZER D.A., LEWIS J. (2011b) poLCA: an R Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software*, Vol. 42, No. 10, 1-29.

[15] MALHOTRA N., DECAUDIN J.-M., BOUGUERRA A. (2007) *Etudes marketing avec SPSS*. Paris, Pearson Education France.

[16] McCUTCHEON A.L. (1987) *Latent class analysis*. Sage University Paper series on Quantitative Applications in the Social Sciences. Beverly Hills and London, Sage Publications.

[17] PAULHUS, D. L. (1991) Measurement and control of response bias. In *Measures of personality and social psychological attitudes* (Vol. 1). San Diego, CA: Academic Press.

[18] QU Y., TAN M., KUTNER M.H. (1996) Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests. *Biometrics*, 52, 797-810.

[19] R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[20] REBOUSSIN B.A., IP E.H., WOLFSON M. (2008) Locally dependent latent class models with covariates: an application to under-age drinking in the USA. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Vol. 171, No. 4, 877-897.

[21] UEBERSAX J.S. (2000) A practical guide to local dependence in latent class models, www.john-uebersax.com/stat/condep.htm.

[22] VAN HERK H., POORTINGA Y.H., VERHALLEN T.M.M. (2004) Response Styles in Rating Scales: Evidence of Method Bias in Data From Six EU Countries. *Journal of Cross-Cultural Psychology*, Vol. 35, No. 3, 346-360.

[23] VERMUNT J.K. (2010) Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. *Political Analysis*, Vol. 18, 450-469.