Part 2. Exploratory Factor Analysis

*we use the brand-attribute rating data to ask the following questions: How many latent factors are there? How do the survey items map to the factors? How are the brands positioned on the factors? What are the respondents’ factor scores?*

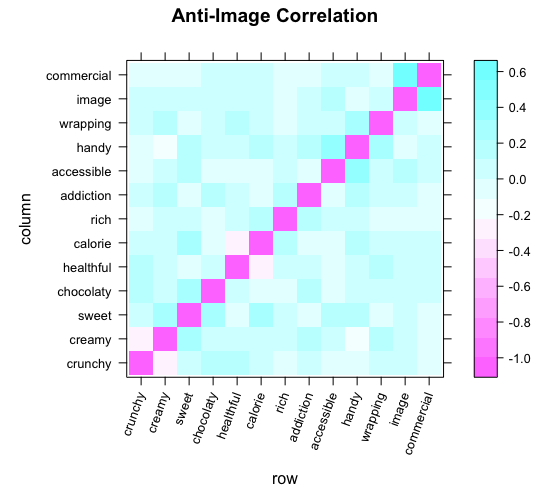
EFA based on the correlation metrics is good at uncovering latent structure and attempt to find a factor structure. It produces results that are very interpretable in terms of the original variables. Furthermore, EFA offers the possibility to check whether the attributes in fact go together in a way that can be interpreted as a single factor, or whether they instead reflect multiple dimensions that we might not have considered

Dimension reduction through the factor analysis is achieved by extracting and synthesizing the overlapping parts among variables in the original dataset into several factors. This requires the correlation among variables not being zero.

**Suitability of factor analysis**

Factor analysis is based on a covariance matrix between variables and assumes that some factors linearly influence the observed model. In other words, the candidate variables (attributes) must have a certain correlation. If there is no correlation between the variables, or the correlation is small, the factor analysis will not be a suitable analysis method. The Kaiser-Meyer-Olkin measure of sampling adequacy (MSA) shows, this dataset is suitable for the degree of factor analysis, since KMO statistic is larger than 0.5. The MSA for individual variables are printed as the diagonal elements of the Anti-image Correlation matrix. The Bartlett’s test for Sphericity compares the correlation matrix to the identity matrix. It checks if there exists relationship between variables that can be summarized with some factors. In this case, it rejects the null hypothesis that the correlation matrix is an identity matrix at 5% level of significance, which indicates the 13 attributes variables are related and therefore suitable for structure detection.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Bartlett's Test** | | | | | | |
| **chisq** | 1235 | **p.value** | 1.4e - 207 | | **df** | 78 |
| **Kaiser-Meyer-Olkin factor adequacy** | | | | | | |
| **Overall MSA = 0.7 MSA for each item:** | | | | | | |
| crunchy | creamy | sweet | chocolaty | healthful | calorie | rich |
| 0.45 | 0.65 | 0.75 | 0.71 | 0.55 | 0.65 | 0.74 |
| addiction | accessible | handy | wrapping | image | commercial |  |
| 0.76 | 0.79 | 0.74 | 0.77 | 0.69 | 0.65 |  |



*Figure.x Factorability check*

**Sample size**

The sample size should be large enough to yield reliable estimates of correlations among the variables. EFA can be reasonably with N/k (Cases / Items) > 5/1. The attribute-brand rating dataset include 13 variables and 500 rows (500/13 = 38).

**The Number of Factors**

We hope that the number of factors should be much smaller than the number of distant variables, while at the same time requiring the retained factors to keep as much information as possible of the original variables. Here we use the eigenvalue method to determine the number of retention factors.

The eigenvalues display the variation that can be explained by the corresponding factors. Factors with an eigenvalue greater than one would be retained. The scree plot visualizes the relationship between the number of factors and the corresponding eigenvalues. 4 components would be selected to do the following factor analysis. On the one hand, the increase in the number of the factors would not bring an increase in the marginal proportion of variance, on the other hand, although 3 factors is suggested, the interpretation power is less than 4 factors.

图片包含 地图

描述已自动生成

**The Rotation of Factors**

The rating variables themselves are ranged from 1 to 5, to normalize the data here is not necessary. Therefore, raw data will be used directly for factor analysis.

In order to interpret the results more easily, rotating is needed. After comparing the oblimin rotation, which allows the dependence among factors, with the orthogonal rotation, which artificially forced the factors to be uncorrelated, I rotate the four-factor solution using orthogonal rotation, since the results is more interpretable. The maximum likelihood approach is employed to extract common factors. The first four factors account for 69 percent of the variance in 13 attributes.

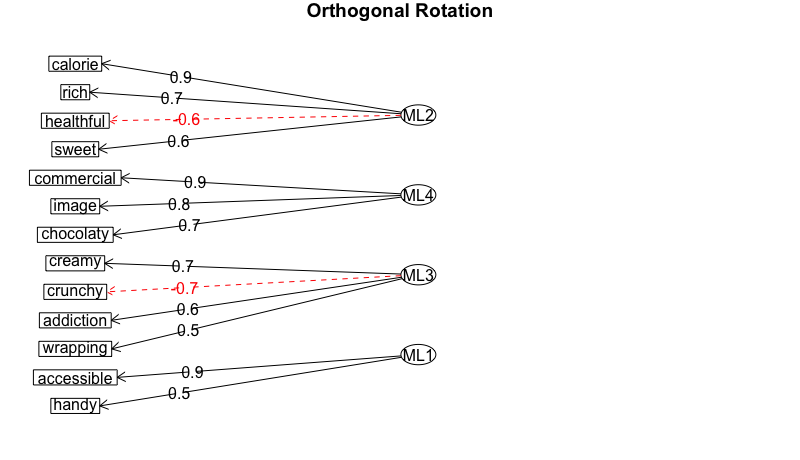
ML2 captures the highest proportion variance (21%) and then followed by ML4 (18%), ML1(15%) and ML3 (15%). Healthful and crunchy are negatively correlated with all the four factors. Loading represents the strength of relationship between a factor and a variable. The first two factors contain the most loadings. Image and commercial are very close to each other in all four factors. If we look at the factor loadings. We can see that calorie, rich and sweet load on the first factor (ML2).

Commercial, image and chocolate load on the second factor (ML4). In most cases, the variability is captured by the four common factors achieve higher than 0.5 (measured by h2).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Factor Analysis using method = Maximum Liklihood** | | | | | | | | |
| rotate = "varimax", scores = "Anderson", max.iter = 1000, fm = "ml" | | | | | | | | |
| Standardized loadings (pattern matrix) based upon correlation matrix | | | | | | | | |
|  | **ML2** | **ML4** | **ML3** | **ML1** | **h2** | **u2** | **com** |
| **crunchy** | -0.19 | -0.01 | -0.27 | 0.55 | 0.55 | 0.446 | 1.5 |
| **creamy** | 0.68 | 0.02 | 0.73 | 1.00 | 1.00 | 0.005 | 2.0 |
| **sweet** | 0.60 | 0.04 | 0.41 | 0.66 | 0.66 | 0.343 | 2.5 |
| **chocolaty** | -0.09 | 0.71 | 0.02 | 0.52 | 0.52 | 0.484 | 1.1 |
| **healthful** | -0.65 | -0.18 | -0.13 | 0.56 | 0.56 | 0.436 | 1.7 |
| **calorie** | 0.92 | 0.11 | -0.27 | 1.00 | 1.00 | 0.005 | 1.4 |
| **rich** | 0.67 | -0.26 | 0.07 | 0.56 | 0.56 | 0.441 | 1.5 |
| **addiction** | -0.07 | 0.17 | 0.6 | 0.4 | 0.40 | 0.603 | 1.2 |
| **accessible** | 0.29 | 0-36 | 0.24 | 1.00 | 1.00 | 0.005 | 1.8 |
| **handy** | 0.15 | 0.3 | 0.08 | 0.33 | 0.33 | 0.667 | 2.0 |
| **wrapping** | -0.16 | -0.2 | 0.48 | 0.52 | 0.52 | 0.478 | 2.6 |
| **image** | 0.16 | 0.79 | 0.21 | 0.89 | 0.89 | 0.108 | 1.8 |
| **commercial** | 0.03 | 0.92 | 0.00 | 1.00 | 1.00 | 0.005 | 1.3 |
|  |  |  |  |  |  |  |  |
|  | **ML2** | **ML4** | **ML3** | **ML1** |  |  |  |
| **SS loadings** | 2.74 | 2.39 | 1.95 | 1.89 |  |  |  |
| **Proportion Var** | 0.21 | 0.18 | 0.15 | 0.15 |  |  |  |
| **Cumulative Var** | 0.21 | 0.39 | 0.54 | **0.69** |  |  |  |
| **Proportion Explained** | 0.31 | 0.27 | 0.22 | 0.21 |  |  |  |
| **Cumulative Proportion** | 0.31 | 0.57 | 0.79 | 1.00 |  |  |  |

**loadings of the factors** on the variables means the relationship of the matrix of factors to the original variables. EFA attempts to find solutions that are maximally interpretable in terms of the manifest variables. In general, it attempts to find solutions in which a small number of loadings for each factor are very high, while other loadings for that factor are low. Different from the Principle Component Analysis (PCA), EFA produces results that are interpretable in terms of the original variables

**Naming the Factors**

****

*Figure.x loadings (|L| > 0.30)*

In order to interpret the factors, let us focus on those attributes with loading > 0.3 by each factor.

• The loadings in ML2 are quite high, which seem very good. Calorie (0.92), rich (0.67) and sweet (0.603) have the highest loading in ML2 (0.92). Healthful (-0.65) locates its highest negative absolute loading also in ML2. We can conclude that ML2 is an **energy factor.**

• In ML4, except accessible, all the other attributes (commercial, image and chocolaty) have their largest loading there. Commercial and image are highly correlated with the advertisement and promotion in products, e.g. where to put the products in the supermarkets, how often and how is advertisement on TV? We can say that ML4 would be an **marketing factor**.

• In ML3, there are a pair opposite tasty attribute: crunchy (-0.67) and creamy (0.73). Wrapping also have a similar score in ML4, which does not contribute too much on explanation power to this factor. Since crunchy or creamy usually describe the filling taste within a chocolate bar, we consider ML3 be a **taste factor.**

ML1 can be seen as a **convenient factor**.

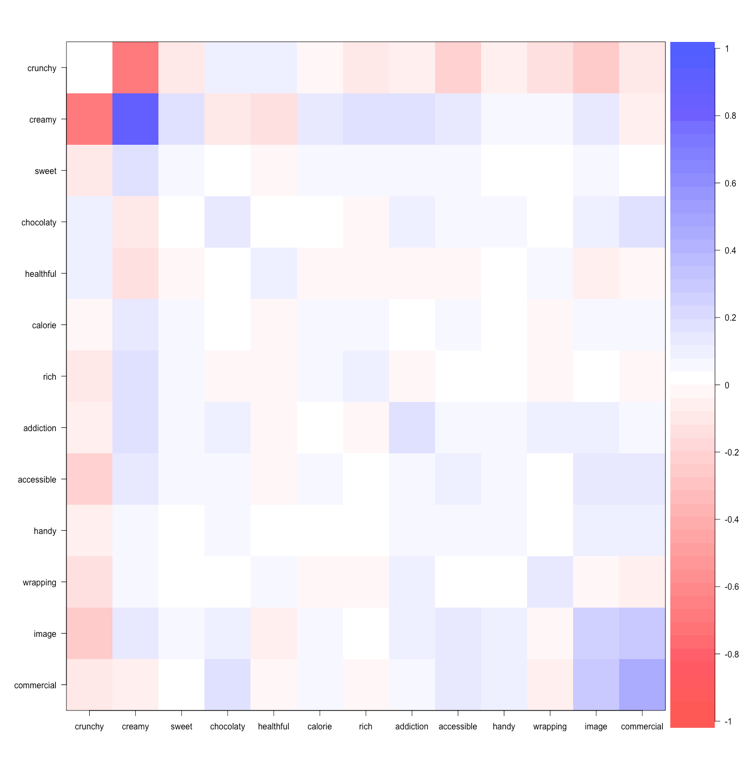
**EFA Scores**

图片包含 屏幕截图

描述已自动生成

The first two factors which highly influence the respondent’s rating are the energy level (e.g. sweet and creamy chocolate bars are more attractive) and the marketing. Consistent with the earlier study, Balisto with the least energy level and has less marketing level, which might lead to respondents evaluate it with a relative low mean rate score. Snickers, Mars and Twix have very high score on both factors. They are similar in such a way: high costumer recognition with rich, sweet and high calorie features, which replenish energy quickly. Most important information has been retained by EFA methods. e.g. Duoplo has the highest commercial EFA score, which is also consistent with the original mean score (4.4/5). Since EFA scores are driven by the correlation matrix, we can also interpret that consumers who like Mars would also like Snicker.

女的打分有关系，男的很多没关系



male <- cov(bars.gender[11:20,3:15])  
cor.plot(male)

