ON ROW AND COLUMN FACTORS

An assignment for Multivariate Statistical Analysis

Student-ID: 616520
Study Subject: Master of Science Business Administration
Statistical Software used: R

Maximilian Suliga
Mainz, 19.03.2021

Table of Contents

Table of Figures

1 The role of row and column factors in Correspondence Analysis and its usage

Factoring multiple rows and columns of categorical data is done in the context of Correspondence Analysis with the major aim to generate a two- or three-dimensional plot, also known as a French or symmetric plot (Bendixen, 1996, p. 26). While the axes of the plot are of smaller importance when it comes to interpreting the plot, observable as well from the fact that it is rarely labeled in applications, it is the row and column factors that actually form both (or all three) axes in these plots. Hence, they are of great importance when it comes to calculating and creating Correspondence Analysis plots but also important when it comes to indicating the relevance of these generated plots, also coined with the term retention (Bendixen, 1996, p. 26).

The starting position for each Correspondence Analysis is a Contingency Table, a table consisting of typically several rows and several columns. Although there exists no formal rule, the rows usually form different individuals and the columns different attributes. In the cells are numbers that indicate how often an attribute is assigned to an individual. As Correspondence Analysis is notably used in Marketing Research, a good example for rows could be product brands of a certain industry and for rows their associated attributes of perception (Kennedy, et al., 1996, pp. 57-60). The number inside a cell could either be the absolute number of positive responses (meaning "this certain attribute describes well the perception of the brand") or alternatively the relative number of total positive responses.

As already indicated above, a Contingency Table requires certain characteristics of the data to be used. To stick with the example above, a marketing research study participant would either mark an attribute for a brand or would not. Hence, all variables need to be binary. The number of participants and therewith the number of positive answers should be reasonable, a rule of thumb could be to have every cell filled out with a predefined minimum number or higher. Other requirements regarding the suitability of Correspondence Analysis for a problem include a too large size of the Contingency Table to be inspected visually alone, homogeneity of the variables and an unknown or poorly understood structure of the data (Lebart, et al., 1984).

While Simple Correspondence Analysis calculates one factor for all rows and for all columns, this can be expanded to more than two total factors using Multiple Correspondence Analysis. For example, for a Multiple Correspondence Analysis of three total factors, either rows or columns can be divided into two groups. Essentially, this “splits up” the factor. For the earlier mentioned example, it would make sense to split the attributes, if they are supposed to describe different groups of perception. Admittedly, a more typical procedure would be to add more variables that belong to a different group, for example the use of the brand measured by asking a participant whether she uses the brand or not.

Another way of describing Correspondence Analysis is viewing it as a deeper analysis of the influence’s origin of a significant Pearson’s chi-squared Test result. The Pearson’s chi-squared test aims to test whether differences in categorical data arose by chance (that means variables are independent) or not. However, the use of this test is limited to the “if”, it does not give an answer to the “why”. Correspondence Analysis decomposes the value of the Chi-squared statistic into two orthogonal components from which dependencies may be analyzed in a more detailed manner (Abdi & Williams, 2010, p. 266).

2 Delimitation of Correspondence Analysis from other Multivariate Statistical Techniques

Correspondence Analysis is an exploratory statistical technique with the aim to explore dependencies of multiple rows and columns at once. The row and column factors needed for this serve to incorporate the maximum of row and column variation. Like Principal Component Analysis or Factor Analysis, Correspondence Analysis ultimately reduces a larger number of variables to a smaller number of factors. While Principial Component Analysis is primarily used for this purpose, Correspondence Analysis puts a bigger stress on the exploratory component of the analysis similar to Factor Analysis. Also, Correspondence Analysis is not intended to be used for predictive models. While Principal Component Analysis maximizes the total variance, it is the chi-squared statistic that is being maximized for Correspondence Analysis (Härdle & Simar, 2015, p. 426). The starting point of the Correspondence Analysis differs as well, as it is a Contingency Table rather than a Covariance or Correlation Matrix (Härdle & Simar, 2015, p. 433). Nevertheless, the factors of both Correspondence Analysis and Principal Component Analysis (denoted as "components") are computed by the Singular Value Decomposition and focus on eigenvalues.

Factor Analysis also differs from its starting point, as it also starts from a Covariance or Correlation matrix. Its main aim in bundling variables however corresponds to that of Correspondence Analysis. Generally, Correspondence Analysis can be viewed as an extension of Factor Analysis, as the main difference lies in the use of metric variables for Factor Analysis and discrete variables for Correspondence Analysis respectively. While discrete variables ultimately drop out for Factor Analysis, metric variables can be transformed to discrete ones using data binning. For example, the metric variable "age" may be binned into the categories "<18", "≥18; <24" etc. (Abdi & Valentin, 2010, p. 366), although it is usually encompassed with a loss of information (Izenman, 2008, p. 634). Further, while Factor Analysis requires a large sample and the assumption of a normal distribution, Correspondence Analysis works equally well for small sample sizes with no distribution assumptions required (Kennedy, et al., 1996, p. 61).

In a broader sense, Correspondence Analysis can also be compared to Canonical Correlation Analysis. Canonical Correlation Analysis aims likewise at finding similar groups of variables by measuring the correlation between sets of variables and can also be used for dimensionality reduction. However, Canonical Correlation Analysis gives an exact value as an output, while the output in Correspondence Analysis is less of a precise nature.

On top of that, Correspondence Analysis may also be used as an approach of clustering a given dataset. Similar to the K-Means algorithm, the ideal result is finding highly homogenous groups that are highly heterogenous to each other. While the final result is given by the K-Means algorithm itself, this is not typically the case for Correspondence Analysis. Rather, the plot that can be created by Correspondence Analysis is used as tool for the human decision maker to group objects together.

3 Computation of row and column factors

There are several different ways in calculating the row and column factors. This work will stick with the approach presented by Härdle & Simar (2015, pp. 428-439). First, the Contingency Table needs to be transformed to a matrix, hereinafter called $\boldsymbol{C}$, with elements

$$c_{ij} = (x_{ij} - E_{ij})E_{ij}^{1/2}, \tag{1}$$

where $E_{ij} = (\sum x_i \sum x_j)/\sum x_{ij}$. Singular Value Decomposition is then applied on $\boldsymbol{C}$**.** In Singular Value Decomposition, a $(J \times K)$ matrix $\boldsymbol{A}$ with $J \leq K$ is decomposed into three components whose products yield the original matrix, $\boldsymbol{A} = \boldsymbol{U\psi V^{\tau}} = \sum_{j}^{J} \boldsymbol{\lambda}_j^{1/2} \boldsymbol{u}_j \boldsymbol{v}_j^{\tau}$ (Izenman, 2008, p. 50). This is done to $\boldsymbol{C}$ $(n \times p)$, yielding

$$\boldsymbol{C} = \boldsymbol{\Gamma\Lambda\Delta}^T. \tag{2}$$

$\boldsymbol{\Gamma}$ contains the eigenvectors of $\boldsymbol{CC}^T$, $\boldsymbol{\Lambda}$ = diag($\boldsymbol{\lambda_1}^{1/2}, \ldots \boldsymbol{\lambda}_R^{1/2}$) with $\boldsymbol{\lambda}_1 \geq \boldsymbol{\lambda}_2 \geq \cdots \geq \boldsymbol{\lambda}_R$ being the eigenvalues of $\boldsymbol{CC}^T$ and $\boldsymbol{\Delta}$ the eigenvectors of $\boldsymbol{C}^T\boldsymbol{C}$. For each computed factor, the eigenvalues express its extracted variance, its sum is referred to as the total inertia of the data matrix (Abdi & Williams, 2010, p. 271). From Singular Value Decomposition, it can be derived that

$$c_{ij} = \sum_{k=1}^{R} \boldsymbol{\lambda}_k^{1/2} \boldsymbol{\gamma}_{ik} \boldsymbol{\delta}_{jk} \ . \tag{3}$$

From equation (3), the two eigenvectors shaping the factors of rows and columns are given in

$$\boldsymbol{\delta}_k = \frac{1}{\sqrt{\lambda_k}} \boldsymbol{C}^T \boldsymbol{\gamma}_k \tag{4}$$

$$\boldsymbol{\gamma}_k = \frac{1}{\sqrt{\lambda_k}} \boldsymbol{C} \boldsymbol{\delta}_k \ . \tag{5}$$

In order to receive the row and column factors, the matrices $\boldsymbol{A}$ $(n \times n)$ and $\boldsymbol{B}$ $(p \times p)$ are taken up with

$$\boldsymbol{A} = \text{diag}\ (\sum x_j) \tag{6}$$

$$\boldsymbol{B} = \text{diag}\ (\sum x_i)\ . \tag{7}$$

Finally, the row and column factors are computed with

$$\boldsymbol{r}_k = \boldsymbol{A}^{-1/2} \boldsymbol{C} \boldsymbol{\delta}_k \tag{8}$$

$$\boldsymbol{c}_k = \boldsymbol{B}^{-1/2} \boldsymbol{C}^T \boldsymbol{\gamma}_k\ , \tag{9}$$

where $\boldsymbol{r}_k$ denotes the vector containing the row factors and $\boldsymbol{c}_k$ the vector with the column factors respectively.

4 Properties of row and column factors

The first and second order moments of the previously computed row and column factors are given by

$$\bar{\boldsymbol{r}}_k = \frac{1}{\Sigma x_{ij}} \boldsymbol{r}_k^T \boldsymbol{a} = 0 \tag{10}$$

$$\bar{\boldsymbol{c}}_k = \frac{1}{\Sigma x_{ij}} \boldsymbol{c}_k^T \boldsymbol{b} = 0 \, , \tag{11}$$

with $\boldsymbol{a} = \boldsymbol{A}1_n$ and $\boldsymbol{b} = \boldsymbol{B}1_p$ respectively, and

$$\mathrm{Var}(\boldsymbol{r}_k) = \frac{1}{\Sigma x_{ij}} \sum_{i=1}^{n} \Sigma x_j \; \boldsymbol{r}_{ki}^2 = \frac{\boldsymbol{r}_k^T \boldsymbol{A} \boldsymbol{r}_k}{\Sigma x_{ij}} = \frac{\lambda_k}{\Sigma x_{ij}} \tag{12}$$

$$\mathrm{Var}(\boldsymbol{c}_k) = \frac{1}{\Sigma x_{ij}} \sum_{j=1}^{p} \Sigma x_i \; \boldsymbol{c}_{kj}^2 = \frac{\boldsymbol{c}_k^T \boldsymbol{B} \boldsymbol{r}_k}{\Sigma x_{ij}} = \frac{\lambda_k}{\Sigma x_{ij}} \tag{13}$$

(Härdle & Simar, 2015, p. 431). Further properties of the row and column factors include

$$\boldsymbol{r}_k^T \, \boldsymbol{A} \boldsymbol{r}_k = \lambda_k \tag{14}$$

$$\boldsymbol{c}_k^T \, \boldsymbol{B} \boldsymbol{c}_k = \lambda_k \tag{15}$$

and

$$\boldsymbol{r}_k = \frac{1}{\sqrt{\lambda_k}} \boldsymbol{A}^{-1/2} \boldsymbol{C} \boldsymbol{B}^{1/2} \boldsymbol{c}_k \tag{16}$$

$$\boldsymbol{c}_k = \frac{1}{\sqrt{\lambda_k}} \boldsymbol{B}^{-1/2} \boldsymbol{C}^T \boldsymbol{A}^{1/2} \boldsymbol{r}_k \, . \tag{17}$$

As composed by the rows and columns of the Contingency Table, the absolute contributions of each row and column are computed by

$$ctr_{i,\ell} = \frac{(\sum_{j=1}^{k} x_i) r_{i,\ell}^2}{\lambda_\ell} \tag{18}$$

$$ctr_{j,\ell} = \frac{(\sum_{i=1}^{n} x_j) c_{j,\ell}^2}{\lambda_\ell} \, , \tag{19}$$

With $r_{i,\ell}$ being the ith element of $\boldsymbol{r}$ for the ℓth eigenvalue and $c_{j,\ell}$ being the jth element of $\boldsymbol{c}$ for the ℓth eigenvalue respectively (Abdi & Valentin, 2010, pp. 369-370). These contribution values take values between 0 and 1 and their sum yields 1. The quality of contribution is measured with the so-called "squared cosine" which is measured by

$$cos_{i,\ell}^2 = \frac{r_{i,\ell}^2}{d_{r,i}^2} \tag{20}$$

$$cos_{j,\ell}^2 = \frac{c_{i,\ell}^2}{d_{c,j}^2} \, , \tag{21}$$

with $d_{r,i}^2$ and $d_{c,j}^2$ being respectively the i-th element of the squared row factors diagonal and the j-th element of the squared column factors diagonal (Abdi & Williams, 2010, p. 271). Essentially, this is the decomposition of the squared distance of an element along the factors.

5 Procedure for variables with more than two attributes

Correspondence Analysis is particularly useful for binary variables and through One Hot Encoding, it is also possible to convert variables with more than two categories into binary categories. However, every new variable increases dimensionality and therefore inflates total inertia (Abdi & Valentin, 2010, p. 370). This results into an underestimation of the percentage of inertia explained by the first dimension. As it is the eigenvalues less or equal to $\frac{1}{K}$, with K being the number of variables, that code the additional dimensions, the formula for computing the corrected eigenvalue can be expressed as:

$$ {}_{c}\lambda_{\ell} = \begin{cases} \left[\left(\frac{K}{K-1}\right)\left(\lambda_{\ell} - \frac{1}{K}\right)\right]^{2} & if\ \lambda_{\ell} > \frac{1}{K} \\ 0 & if\ \lambda_{\ell} \leq \frac{1}{K} \end{cases} . \tag{18} $$

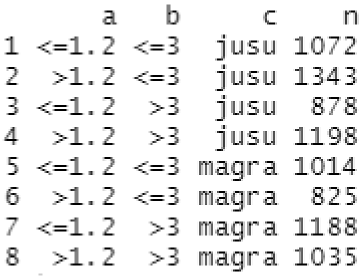
Alternatively, the percentage of inertia relative to the average inertia of the off-diagonal blocks of the “Burt matrix” can be used, where the “Burt matrix” is standing for the $J \times J$ table $\boldsymbol{B}$ obtained from $\boldsymbol{B} = \boldsymbol{C}^{T}\boldsymbol{C}$ (Abdi & Valentin, 2010, p. 369). This yields to the following equation, with

$$ \bar{\mathcal{J}} = \frac{K}{K-1}\left(\sum\nolimits_{h} \lambda_{\ell}^{2} - \frac{J-K}{K^{2}}\right) , \tag{19} $$

where $\bar{\mathcal{J}}$ stands for the average inertia (Abdi & Valentin, 2010, pp. 370-372).

6 Application, Interpretation and Visualization

The following application is based on artificially generated data. The code used for application in R is mostly based on Kassambara, 2017 (pp. 83-106) and can be found in the appendix. Tables and figures included show the exact output given by RStudio when running the code. The data used needs to be prepared before the Correspondence Analysis functions can work with it. Table 1 shows the structure of the data before input. The columns a, b and c are qualitative binary variables, its possible characteristics are shown below the respective column name. The column n gives the frequency for each row. There are eight rows which depict each possible combination of the column variables. The row numbers can hence be seen as individuals with each being unique.



```
       a   b     c    n
1 <=1.2 <=3  jusu 1072
2  >1.2 <=3  jusu 1343
3 <=1.2  >3  jusu  878
4  >1.2  >3  jusu 1198
5 <=1.2 <=3 magra 1014
6  >1.2 <=3 magra  825
7 <=1.2  >3 magra 1188
8  >1.2  >3 magra 1035
```

Table 1 snapshot of the data after first preparation

Conducting Multiple Correspondence Analysis to the data, two Factor Maps are given represented in Figure 1 & 2 as well as a plot of the Variables representation in Figure 3. All three figures are plots using the two row and column factors with the greatest explained Variance of the table as axes. As these factors are simultaniously called dimensions in literature (Abdi & Williams, 2010, p. 366), the output given by R labels them as dimensions as well. The Factor Maps facilitate separate analysis of the variables and individuals respectively, this can be of particlar interest when the plot containing both is overstocked. This does not apply for the given data and therefore further analysis will continue with the following French plot.

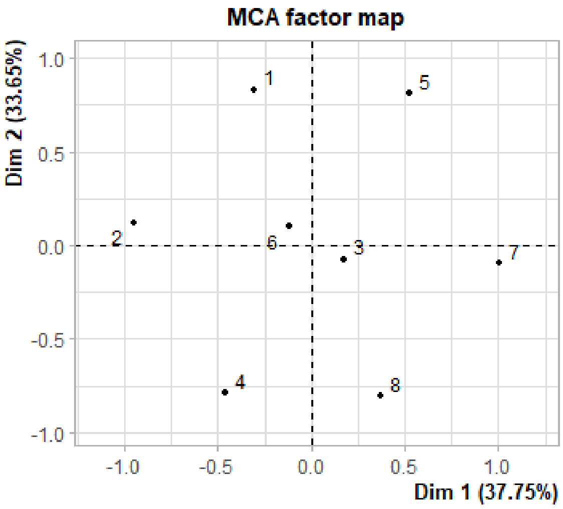


Figure 1 Factor Map of the rows

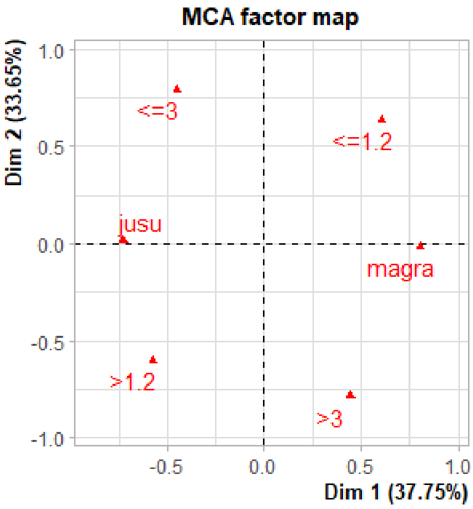


Figure 2 Factor Map of the columns

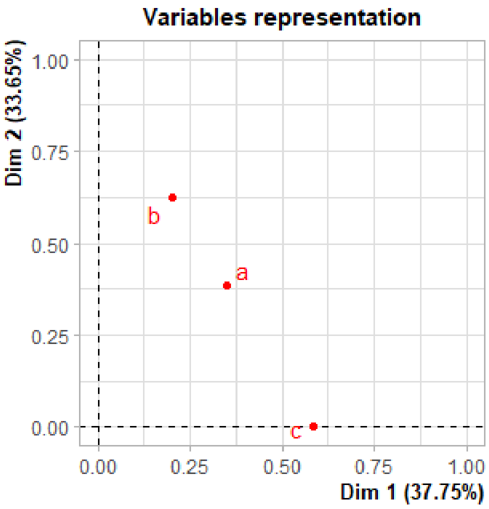
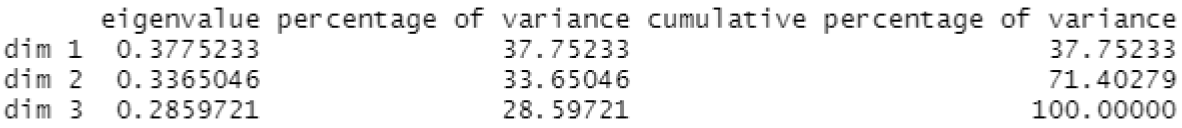


Figure 3 representation of columns by each Factor

Figure 3 shows the represetation of the variables by each plot. While it may not be of particular interest for variables a and b, c is standing out as being virtually unexplained by the second factor. This is also visible in Figure 2, where c's characteristics "jusu" and "magra" lie nearly on the first axis.

The eigenvalues as well as the percentage amount of each factor explained can be extracted and visualized seperately. In this example, the eigenvalue equals the total variance. This is a rather unlikely occurrence, as it does not appear in examples provided by literature, as for instance in Härdle & Simar (2015, p. 433). In this example, it comes from the sum of the eigenvalues being 1. It is also characteristic that all factors vary only by little, which is not the case in classic demonstrations, like in Izenman (2008, p. 646). These variances amount to the total inertia, which can be read from the last column of Table 2.



```
      eigenvalue percentage of variance cumulative percentage of variance
dim 1  0.3775233               37.75233                          37.75233
dim 2  0.3365046               33.65046                          71.40279
dim 3  0.2859721               28.59721                         100.00000
```

Table 2 eigenvalues and percentage of explained variance

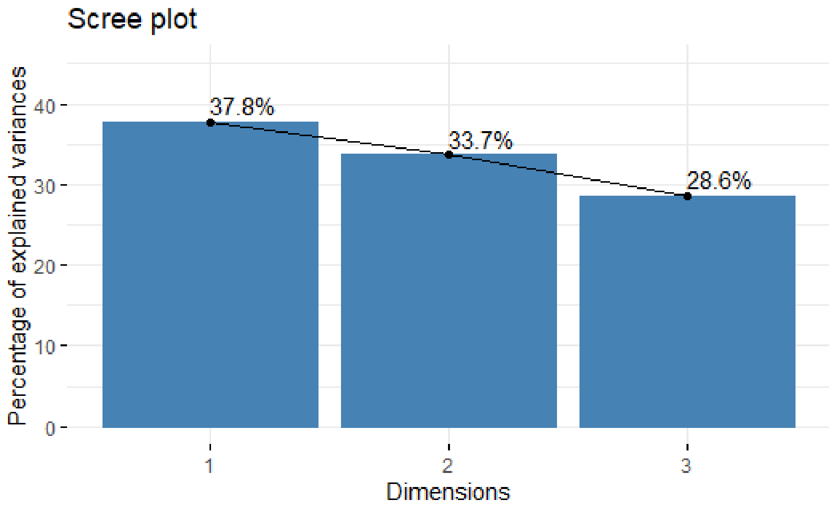


Figure 4 explained variance for each Factor

For the two-dimensional French plot, the two factors with the largest variance are used as dimensions and hence the retention for the plot equals 71.4%. Correspondence Analysis can be seen as a trade-off between dimensionality and unexplained variance. A contingency table with two variables does not need any reduction as it can be plotted in an easily interpretable two-dimensional plot. A contingency table with three variables is still possible to visualize, although harder to interpret and with more than three dimensions it becomes impossible to get any picture of the table for interpretation. As taken from the example by Bendixen (1996, p. 22), a 15-dimensional space may be a perfect graphical representation for a contingency table with 15 variables, as its retention equals 100%. However, it is much more practicable to work with a two-dimensional space and hence a retention rate of perhaps 75% may be considered as a beneficial trade-off.

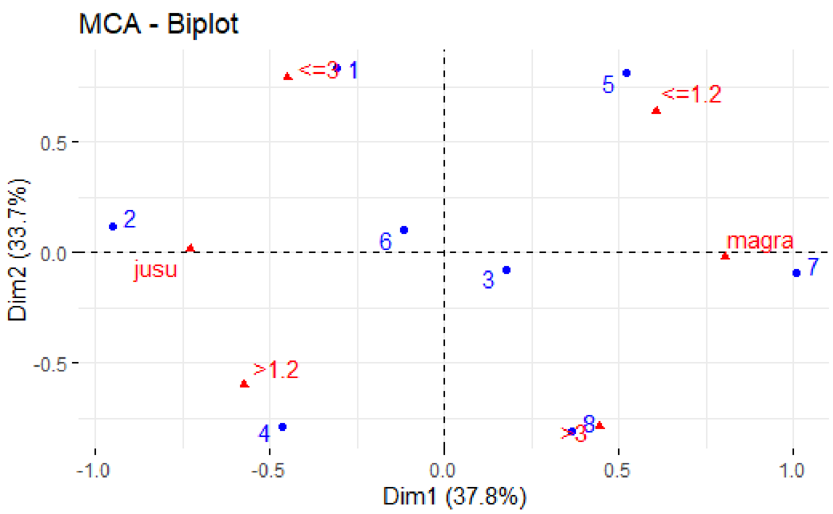
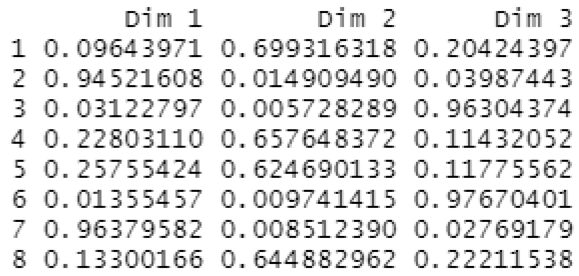


Figure 5 French plot combining both Factor Maps

Next, the actual French plot is viewed in Figure 5, denoted as "Biplot" by R. Essentially, this plot is a merge of both Factor Maps. Rows and columns are both quite spread out overall, forming a circle-like shape. Individuals 3 and 6 stick out as they lie both significantly closer to the origin of the plot. This means that their scores for each row is close to the average score of each row, resulting into an overall low variance and hence a short distance away from the origin (Härdle & Simar, 2015, p. 432). Individuals that lie close to each other share a similar profile and may therefore be grouped together as part of a Clustering approach. For this example, this could only be applied to rows 3 and 6 which both have a profile similar to the average, all other individuals seem to be equally and far enough apart to leave groupings out. The same applies to the variables, which are all similarly spread out across the map. Any groupings for variables can be regarded as questionable in this specific case.

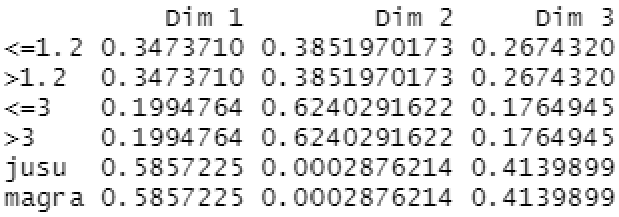
Combining rows and columns, individuals with a high weight in a column lie closer to the respective column, while individuals with a low weight in a column lie far apart from it in the plot (and vice versa). In this example, every characteristic seems to have one individual with a particularly high weight, with individual 8 and characteristic >3 depicting the closest relationship. Put in another way, one could say that there are six individuals that are clearly defined by a certain characteristic assigned to them and two individuals that represent profiles with a bit of all. A proper example could be eight study profiles of a university faculty with six classes from different chairs. Six of these study profiles can be assigned clearly to each chair and can therefore be called degrees with a major in the respective chair, while the other two can be labeled as broad degrees with no particular major.

As observed from the upper plots, some points are better displayed than others by the two largest row and column factors. This quality can be measured using the squared cosines (Abdi & Williams, 2010, pp. 272-273) and these are given for each variable and individual of our example by the tables below. As can be seen from Table 3, Individuals 2 and 7 score particularly high for the first factor, which can also be observed when looking precisely at the distance from the first axis. Individuals 2, 3, 6 and 7 score particularly low for the second factor. This is also no surprise as they lie close to the first axis. While individuals 2 and 7 score particularly low for Factor 3, individuals 3 and 6 score particularly high. Since the third factor is containing all remaining variance, this easily explains the total distance of the points to the origin in the French plot. Individuals 2 and 7 are barely explained by Factor 3, hence they are mostly explained by the remaining two factors. Looking closely at the plot, their distance to the origin is confirmatory the largest. On the other hand, individuals 3 and 6 are mostly explained by Factor 3, therefore they are not explained by a lot by the remaining Factors and hence their distance to the origin is unarguably the lowest.



```
       Dim 1       Dim 2      Dim 3
1 0.09643971 0.699316318 0.20424397
2 0.94521608 0.014909490 0.03987443
3 0.03122797 0.005728289 0.96304374
4 0.22803110 0.657648372 0.11432052
5 0.25755424 0.624690133 0.11775562
6 0.01355457 0.009741415 0.97670401
7 0.96379582 0.008512390 0.02769179
8 0.13300166 0.644882962 0.22211538
```

Table 3 squared cosines for the rows



```
          Dim 1        Dim 2     Dim 3
<=1.2 0.3473710 0.3851970173 0.2674320
>1.2  0.3473710 0.3851970173 0.2674320
<=3   0.1994764 0.6240291622 0.1764945
>3    0.1994764 0.6240291622 0.1764945
jusu  0.5857225 0.0002876214 0.4139899
magra 0.5857225 0.0002876214 0.4139899
```

Table 4 squared cosines for the columns

In Table 4, the categories jusu and magra score particularly low for the second factor. This is no surprise, as these two lie almost on the second axis of the French plot. Logically, the remaining variance has to be in the third factor and therefore both categories are the ones scoring the highest for the third factor. Another way of measuring the quality of association of a row or column to a Factor is by inspecting its squared cosine. Figures 6 and 7 visualize the squared cosines of the total Factors.

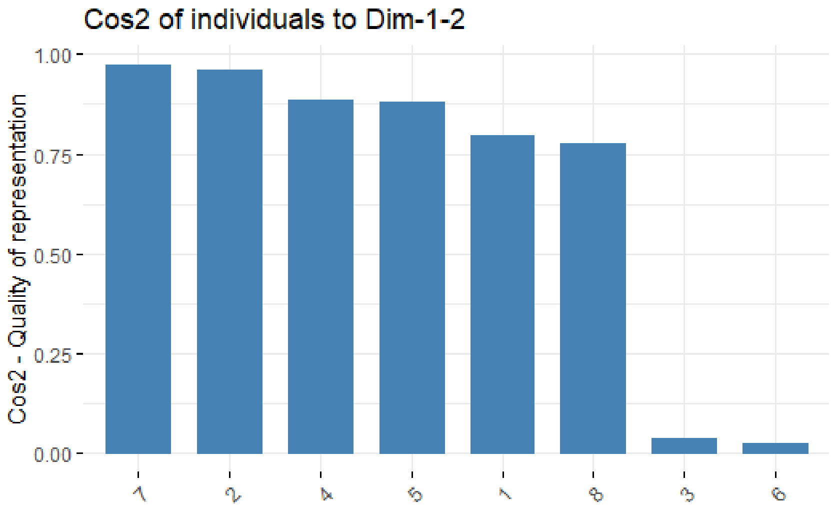


Figure 6 squared cosines of each row for both Factors

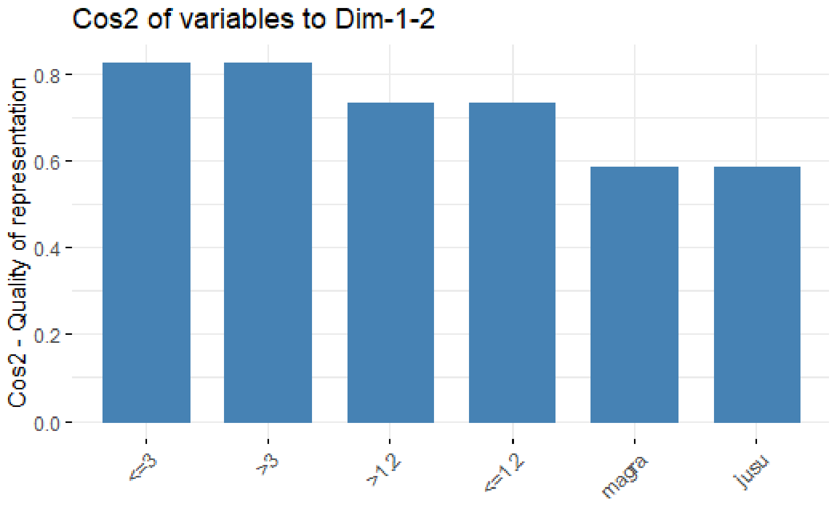
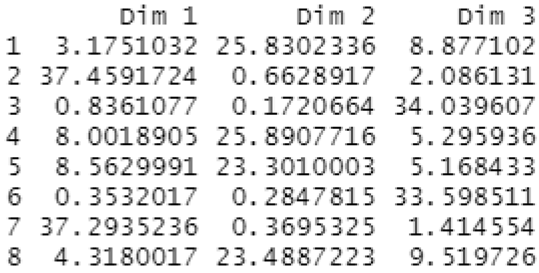


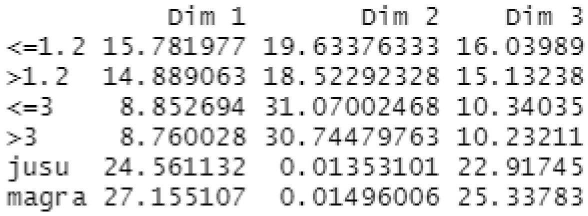
Figure 7 squared cosines of each column for both Factors

Lastly, contributions to row and column factors are analyzed. The absolute contribution in percent for each column and row is given by Table 5 and 6. The patterns are similar to the squared cosines. In order to detect potential outliers, both contribution as well coordinates of each row and column need to be considered. The coordinates denote the number of standard deviations a row or column is away from the origin in the respective dimension. Outliers are typified by being more than one standard deviation away from the origin and contributing above average to the respective factor (Bendixen, 1996, p. 33).



```
       Dim 1      Dim 2     Dim 3
1  3.1751032 25.8302336  8.877102
2 37.4591724  0.6628917  2.086131
3  0.8361077  0.1720664 34.039607
4  8.0018905 25.8907716  5.295936
5  8.5629991 23.3010003  5.168433
6  0.3532017  0.2847815 33.598511
7 37.2935236  0.3695325  1.414554
8  4.3180017 23.4887223  9.519726
```

Table 5 absolute contribution of every row to each Factor



```
          Dim 1       Dim 2    Dim 3
<=1.2 15.781977 19.63376333 16.03989
>1.2  14.889063 18.52292328 15.13238
<=3    8.852694 31.07002468 10.34035
>3     8.760028 30.74479763 10.23211
jusu  24.561132  0.01353101 22.91745
magra 27.155107  0.01496006 25.33783
```

Table 6 absolute contribution of every column to each Factor

Detecting above-average contribution can be easily done by visualizing the contribution of all rows and columns on all Factors by plotting them. In Figures 10-13, each bar shows the respective contribution and each red dotted line denotes the respective average contribution. Obviously, there always needs to be at least one row or column contributing above average as long as contribution is not constant. Hence, only Individuals 3 and 6 can be excluded already by this point.

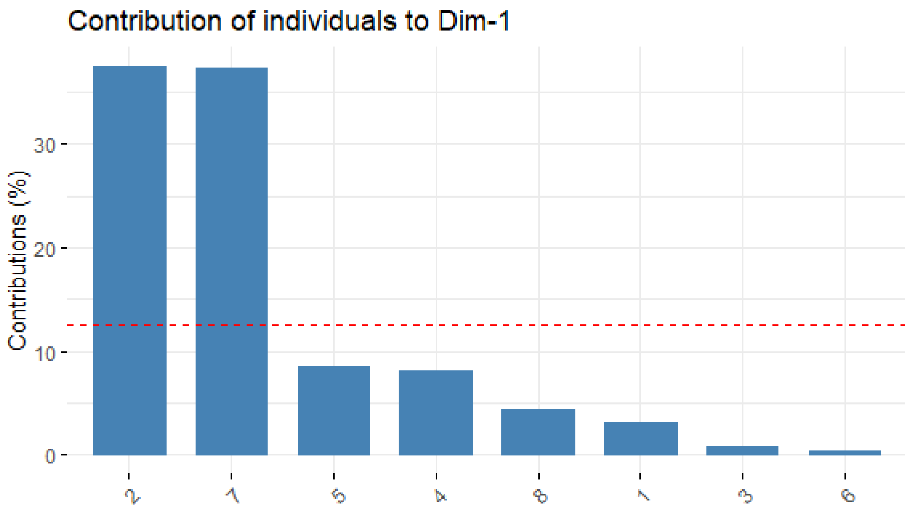


Figure 8 contribution of each row to Factor 1

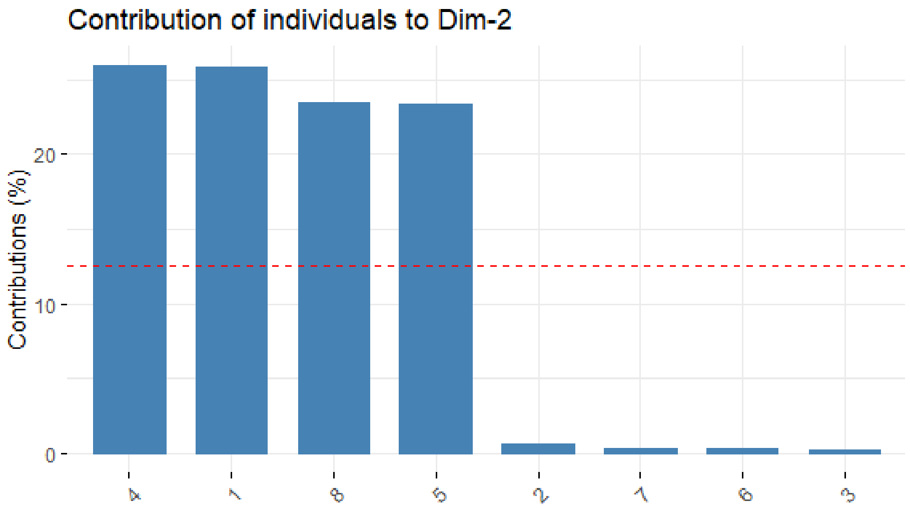


Figure 9 contribution of each row to Factor 2

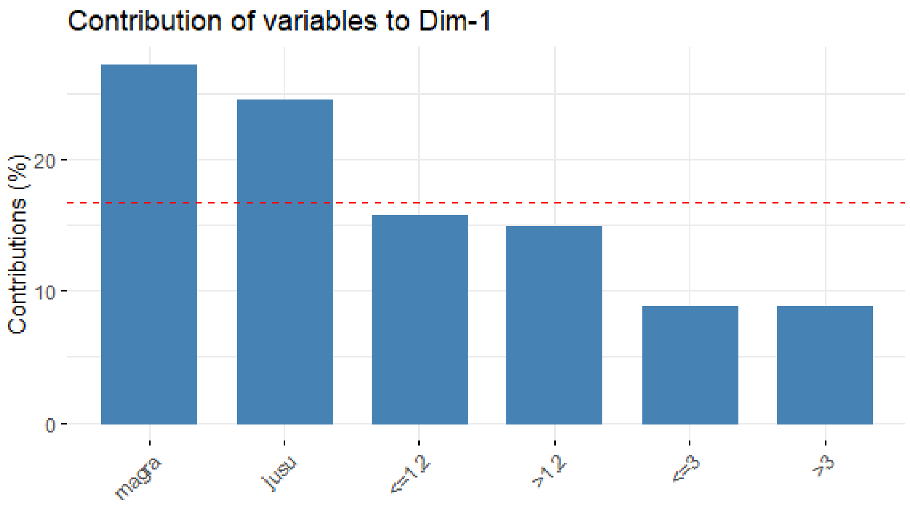


Figure 10 contribution of each column to Factor 1

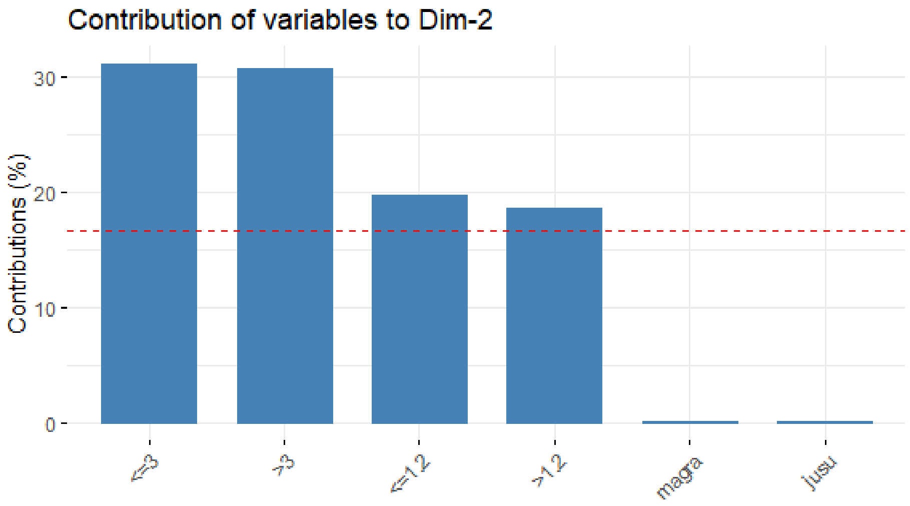
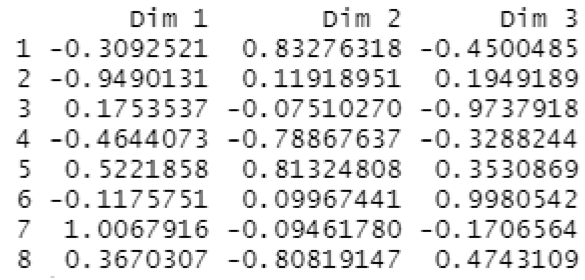


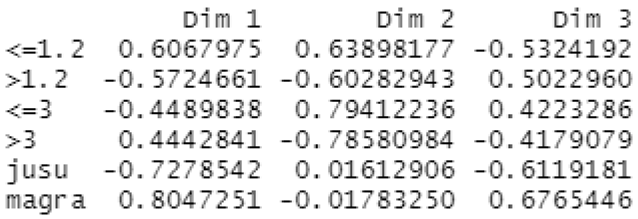
Figure 11 contribution of each column to Factor 2

In the next step, the coordinates are inspected in Table 7 and 8. From the coordinates, only Individual 7 is slightly above 1 for Factor 2. Individual 7 may therefore be regarded as a potential outlier. However, due to the unusually small number of rows and columns and the fact that the coordinate for the second Factor of Individual 7 barely exceeds 1, dropping Individual 7 may be disregarded for this example.



```
        Dim 1       Dim 2      Dim 3
1 -0.3092521  0.83276318 -0.4500485
2 -0.9490131  0.11918951  0.1949189
3  0.1753537 -0.07510270 -0.9737918
4 -0.4644073 -0.78867637 -0.3288244
5  0.5221858  0.81324808  0.3530869
6 -0.1175751  0.09967441  0.9980542
7  1.0067916 -0.09461780 -0.1706564
8  0.3670307 -0.80819147  0.4743109
```

Table 7 row coordinates for each Factor



```
            Dim 1       Dim 2      Dim 3
<=1.2  0.6067975  0.63898177 -0.5324192
>1.2  -0.5724661 -0.60282943  0.5022960
<=3   -0.4489838  0.79412236  0.4223286
>3     0.4442841 -0.78580984 -0.4179079
jusu  -0.7278542  0.01612906 -0.6119181
magra  0.8047251 -0.01783250  0.6765446
```

Table 8 column coordinates for each Factor

7 Conclusion

A rather neglected technique of Multivariate Statistics, Correspondence Analysis is incredibly suitable for survey data and complements Principal Component Analysis and Factor Analysis by being aimed at categorical variables. It incorporates Multivariate Statistics’ exploring, explaining and grouping parts through visualizing the row and column factors of its input Contingency Table. This makes it a technique that can be compared in various kinds to other Multivariate Statistical Techniques. The recognition of Correspondence Analysis therefore deserves more attention in theory as well as in practice.

Not observable from the software in use, it is the calculation of row and columns factors that play a crucial part for Correspondence Analysis and its most important output, the French plot. They capture the variation of the categorical variables and enable the de-dimensioning of the data, which in turn makes a large dataset of many variables both able to be visualized as well as to be interpreted. Correspondingly, extended analysis in identifying potential outliers also relies heavily on them. It is therefore mandatory to know their role and properties for conducting Correspondence Analysis properly.

References

Abdi, H. & Valentin, D., 2010. Multiple Correspondence Analysis. In: J. Gower , S. Lubbe & N. le Roux, Hrsg. *Understanding Biplots.* New York: John Wiley & Sons, pp. 365-403.

Abdi, H. & Williams, L. J., 2010. Correspondence Analysis. In: N. J. Salkind, Hrsg. *Encyclopedia of Research Design.* s.l.:Sage, pp. 267-278.

Bendixen, M. T., 1996. A Practical Guide to the Use of Correspondence Analysis in Marketing Research. *Marketing Research On-Line,* Band I, pp. 16-38.

Härdle, W. K. & Simar, L., 2015. *Applied Multivariate Statistical Analysis.* 4. Hrsg. Berlin, Heidelberg: Springer.

Izenman, A. J., 2008. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* 1st Hrsg. New York: Springer.

Kassambara, A., 2017. *Practical Guide To Principal Component Methods in R.* 1. Hrsg. s.l.:STHDA.

Kennedy, R., Riquier, C. & Sharp, B., 1996. Practical Applications of Correspondence Analysis to Categorical Data in Market Research. *Journal of Targeting Measurement and Analysis for Marketing,* V(1), pp. 56-70.

Appendix

```
#Eventually the install.packages(“”) function will be needed.
library("FactoMineR")
library("factoextra")

# Custom function which converts from a data frame of counts to a data frame of cases as this is needed as an input for the MCA function.
countsToCases = function(x, countcol = "Freq") {
  # Get the row indices to pull from x.
  idx <- rep.int(seq_len(nrow(x)), x[[countcol]])

  # Drop count column.
  x[[countcol]] = NULL

  # Get the rows from x.
  x[idx, ]
}

df = read.csv("MaximillianSuliga.csv", header=TRUE)

#Generate the sums of n for each categorical observation. Can be also used for the MCA function with some adjustments.
dff = aggregate(n~ a+b+c, data=df, FUN = "sum")
#A "default" input for the MCA function.Every possible observation is stores seperately rather than being summed up.
```

```
dff.long = countsToCases(dff, countcol = "n")


dff


mca1 = MCA(dff[,-4], row.w = dff[,4], graph = TRUE)

#First, the frequency column n is dropped, leaving only the categorical data. With "row.w", the categorical data is being weighted by the respective n-value. This yields to exactly 8 Individuals, as all rows with the same attributes are summed together.

mca2 = MCA(dff.long, graph = TRUE)

#Every observation is being fed as one column, creating a very long list of observations


mca1

#As the large amount of principally equal observations makes the mca2 plots harder to read,analysis continues with mca1 only. the results are exactly the same as an be observed by the plots, only readability has improved compared to mca2.


#The command for extracting the eigenvalues as well as the explained variance.

mca1$eig


#Both summary commands for rows and columns respectively. This analysis will rather analyze property by property than all properties for rows followed by all properties for columns.

mca1$ind

mca1$var


#To avoid eventual errors, the results for the rows and the columns are extracted into these variables.

ind = get_mca_ind(mca1)
```

```
var = get_mca_var(mca1)


#Plot of the dimensions and their explained variance.

fviz_screeplot(mca1, addlabels = TRUE, ylim = c(0, 45))


fviz_mca_biplot(mca1,

          repel = TRUE, # Avoid text overlapping.

          ggtheme = theme_minimal()) #French plot of the data.


#Extraction of the squared cosine values for rows and columns.

ind$cos2

var$cos2


#Plot of the squared cosine values for rows and columns.

fviz_cos2(mca1, choice = "ind", axes = 1:2)

fviz_cos2(mca1, choice = "var", axes = 1:2)


#The following two commands are added in the appendix, but their output is not in the text of the paper. Their output had been generated properly and stored in the paper file. For unknown reasons, the word processing program has crashed and most of the content including plots was gone. In an attempt to restore the file, the computer with which this work has been accomplished was restarted. While most of the text could be restored, all plots were immutably gone. As they were generated again in RStudio with the exact same code as before, the output of these commands resulted to different unusable outcomes. According to the official website of R, it can happen occasionally that plots are not shown properly due to various reasons. Solution proposals have not succussed.

fviz_mca_var(mca1, col.ind = "cos2",

        gradient.cols = c("white", "blue", "red"),
        repel = TRUE, # Avoid text overlapping
        ggtheme = theme_minimal())

fviz_mca_ind(mca1, col.var = "cos2",
        gradient.cols = c("white ", " blue ", " red "),
        repel = TRUE, # Avoid text overlapping
        ggtheme = theme_minimal())

# Contributions of rows to dimension 1 and 2 and of columns to dimension 1 and 2.
fviz_contrib(mca1, choice = "ind", axes = 1, top = 15)
fviz_contrib(mca1, choice = "ind", axes = 2, top = 15)
fviz_contrib(mca1, choice = "var", axes = 1, top = 15)
fviz_contrib(mca1, choice = "var", axes = 2, top = 15)

#Contribution of row and column factors.
ind$contrib
var$contrib

#Coordinates of row and column factors.
ind$coord
var$coord
```

Declaration

I hereby confirm that I have authored this document independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

I have attached the code used to produce the analysis in the appendix. I confirm that I have written and executed the analysis, and that the code is complete and executable.

MAINZ, 19.03.2021

Max [illegible]