CHAPTER 19

Exemplars, categories, prototypes

Simple and multiple correspondence analysis

What you will learn from this chapter:

This chapter introduces Correspondence Analysis. It is similar to PCA, but is designed for visualization and exploration of bivariate and multivariate categorical data. The first case study focuses on register variation of English Basic Colour Terms by using Simple Correspondence Analysis, which can be used for visualization of bivariate categorical data in two-dimensional contingency tables. In the second case study of German lexical categories *Stuhl* 'chair' and *Sessel* 'armchair', you will learn how to perform Multiple Correspondence Analysis with higher-dimensional tables.

19.1 Register variation of Basic Colour Terms: Simple Correspondence Analysis

19.1.1 The data and hypothesis

To reproduce the code in this case study, you will need several add-on packages. These packages need to be installed and loaded, if you have not done so previously.

```
> install.packages(c("vcd", "ca", "rgl"))
> library(Rling); library(vcd); library(ca); library(rgl)
```

The dataset for this case study was introduced in Chapter 4, which explored the dispersion of Basic Colour Terms (BCT) in a corpus. The dataset is called `colreg`. It contains the counts of eleven BCT in different registers from the Corpus of Contemporary American English (Davies 2008 –): spoken data on television and radio, fiction, academic prose and press (newspapers and magazines combined).

```
> data(colreg)
> colreg
```

```
        spoken   fiction   academic   press
black   20335    41118     26892      73080
blue    4693     22093     3605       21210
brown   1185     10914     1201       11539
gray    1168     12140     1289       6559
green   3860     14398     4477       26837
orange  931      3496      474        5766
pink    962      7312      584        6356
purple  613      3366      429        3403
red     7230     25111     5621       34596
white   14474    40745     26336      54883
yellow  1349     10553     1855       10382
```

The case study in Chapter 4 revealed that the frequencies of primary and secondary BCT are distributed across the registers unequally. More specifically, the secondary terms (*brown*, *gray*, *orange*, *pink* and *purple*) are less evenly distributed in the corpus than the primary terms (*black*, *white*, *red*, *green*, *yellow* and *blue*). The present case study will explore which terms are attracted to which register. A popular tool for visualization of categorical data is the mosaic plot:[1]

```
> mosaicplot(colreg, las = 2, shade = TRUE, main = "Register
variation of BCT")
```

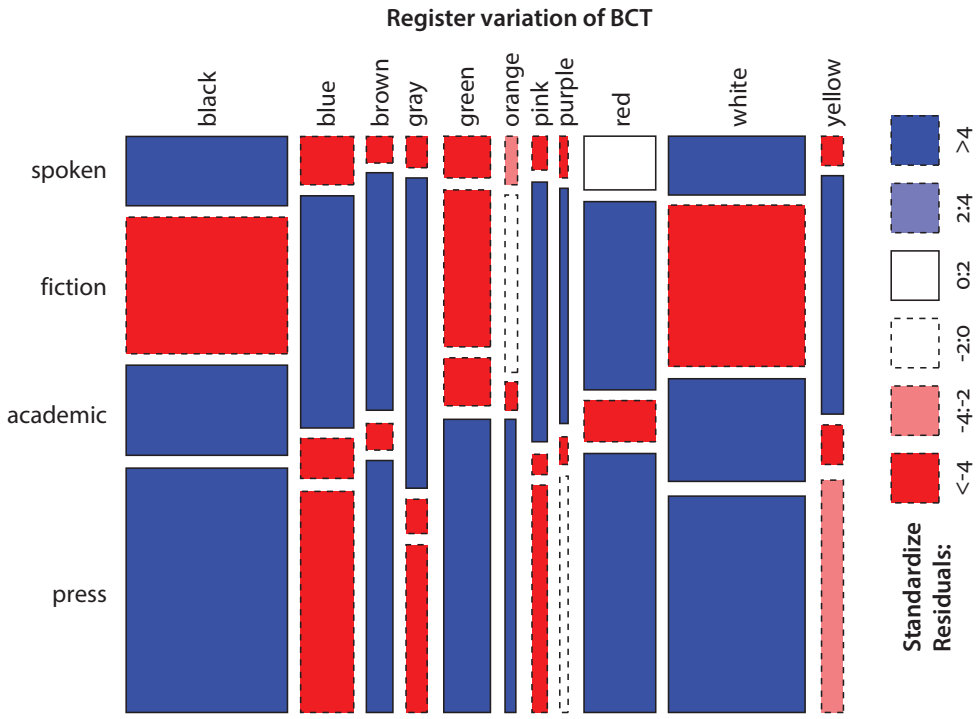

Figure 19.1. Mosaic plot of register variation of BCT

1. This function is similar to *mosaic()* in vcd (see Chapter 9), which offers more graphical and structural options.

The result is shown in Figure 19.1. It shows which colour terms are overrepresented in a given register (blue-shaded rectangles) and which ones are underrepresented (pink- and red-shaded rectangles). See Chapter 9 for more information about how mosaic plots are interpreted. However, the mosaic plot is not particularly convenient when the number of categories is large. Moreover, it does not show any common dimensions of variation. A more appropriate method in this situation is Simple Correspondence Analysis (SCA).

19.1.2 Simple Correspondence Analysis

Correspondence Analysis (CA) is useful for identification of systematic relationships between variables and capturing the main tendencies in several dimensions. Similar to MDS, PCA and FA, it represents the objects of analysis as points in a low-dimensional space. We will use an implementation of SCA in the `ca` package. The code is very simple:

```
> ca.bc <- ca(colreg)
```

The details about the CA model can be obtained by using `summary()`. Of particular relevance is the upper part of the output, namely, the table with principal inertias, which show how much variation is explained by each dimension. Principal inertias are CA equivalents of eigenvalues in PCA, which was discussed in the previous chapter:

```
> summary(ca.bc)

Principal inertias (eigenvalues):

dim     value      %     cum%  scree plot
1       0.043730  77.9  77.9  *************************
2       0.010787  19.2  97.1  *****
3       0.001650  2.9   100.0
        ----      ---
Total:  0.056167  100.0
[output omitted]
```

The algorithm tries to represent the associations between variables in as few dimensions as possible. The two first dimensions represent together 97.1% of variation. This is a good approximation. Three dimensions are sufficient to capture all variation, without any loss of information.

What exactly are the associations between the registers and the colour terms like? To answer this question, one can explore a two-dimensional CA map (Figure 19.2):

```
> plot(ca.bc)
```

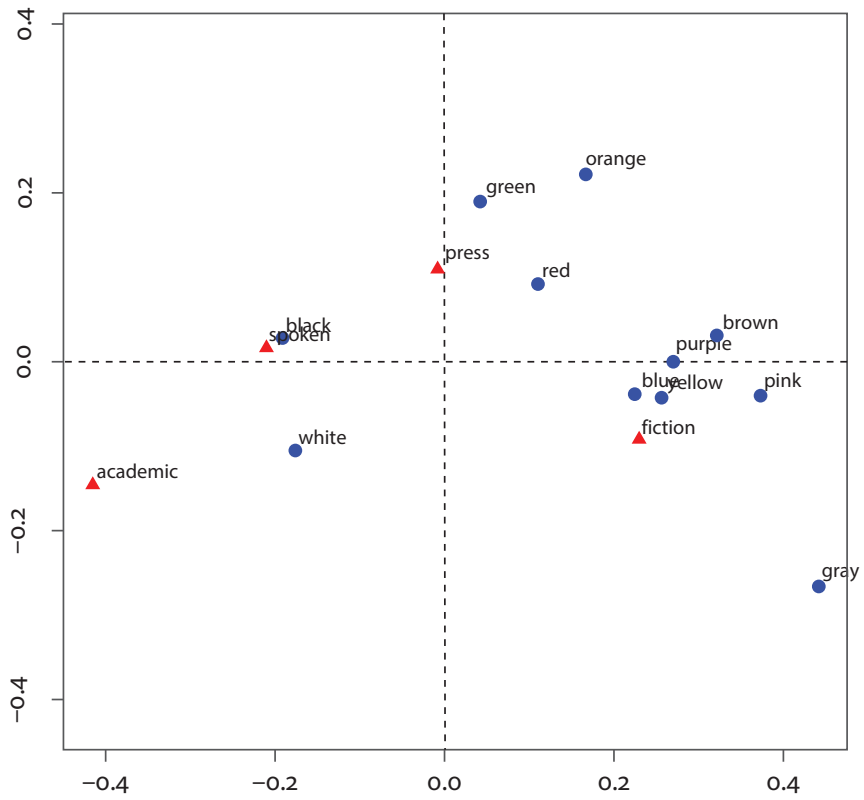

Figure 19.2. Associations between basic colour terms and registers in the Corpus of Contemporary American English (COCA). A two-dimensional correspondence analysis plot

CA plots display the labels of the levels of input variables in a low-dimensional space. In SCA, row labels are located close to one another if they contain similar proportions of counts in each column – in other words, if the rows have similar **profiles**. For example, the profile of *black* in `colreg` is [0.13, 0.25, 0.17, 0.45], and the profile of *white* is [0.11, 0.30, 0.19, 0.40]. The numbers are the actual frequencies divided by the row totals. Those profiles are more similar to each other than to the profile of *gray* [0.06, 0.57, 0.06, 0.31]. This explains why *black* and *white* are close on the map, and *gray* is far from both of them. CA maps represent the difference between profiles as the χ^2-distance. It is similar to the Euclidean distance, but is weighted by the inverse of the corresponding value in the average row profile. As a result, the stronger a row deviates from the average profile, the farther away from the other rows it will be located. The same holds for columns. Their labels are located close if they contain similar proportions of counts in each row, i.e. their profiles are similar. However, the interpretation of mutual proximity of rows (i.e. the BCT) and columns (i.e. the registers) is not straightforward. By default, the function creates a so-called symmetric plot. This means that the algorithm tries to overlay the BCT space on the register space in an optimal way. As a result of the rescaling, the distances between rows and columns are no longer meaningful. Therefore, the location of individual rows (BCT) should be interpreted with regard to the dimensions

formed by the columns (i.e. the registers), rather than on the basis of their proximities with the individual columns.

Interpretation of symmetric (default) CA maps

It is easy to misinterpret a CA map. To be on the safe side, follow these rules:

- Row-to-row distances on the CA map represent the approximate χ^2-distances between the row profiles.
- Column-to-column distances on the CA map represent the approximate χ^2-distances between the column profiles.
- There is no direct interpretation of row-to-column or column-to-row distances. Interpret the dimensions first, and then examine how the profiles are located with regard to the dimensions of variation (Greenacre 2007: 72).

It seems that the first dimension (the horizontal axis) contrasts the achromatic primary colours (black and white), which are located in the left-hand part of the graph, with the other terms, which can be found in the right-hand area. This part of the plot also contains the labels of the spoken and academic subcorpora. In the right-hand part, one can find the label of fiction and those of the secondary colour terms. This is not surprising, since fiction writers tend to use more elaborate and varied attributes for objects than those one can expect to find in the other registers. Interestingly, the primary colours *yellow* and *blue* are also close to the secondary BCTs. The press subcorpus has an intermediary position with regard to the horizontal dimension and is located higher on the vertical dimension than the other registers. This orientation is shared by *red* and *green*, partly because of the political connotations of those colour terms (*Green Party*, *Red Army*), partly due to such proper names as *Red Cross* and *Green Bay Packers*, and partly due to food terms (red wine, green beans) in recipes and articles on nutrition. The secondary term *orange* is also found nearby. A closer inspection suggests a simple explanation: *orange* is frequently used in the sense 'made of oranges'. This use is frequent in recipes and articles on nutrition in the magazines.

To see if there are any interesting patterns that are missed when one looks at two dimensions only, it is possible to plot all three dimensions with the help of the `plot3d()` function in the same package. Although this is not really necessary for our analysis, since

the two-dimensional solution captures nearly all variance, a three-dimensional solution will be explored for purposes of illustration. Note that you also have to install and load another package, `rgl`, to create three-dimensional interactive plots (see also Section 17.3 in Chapter 17).

```
> plot3d(ca.bc, labels = c(1,1))
```

The argument `labels = c(1, 1)` means that both row and column profiles should be shown as text labels, while `c(0, 0)` will produce only symbols on the plot, and `c(2, 2)` will show a combination of symbols and labels. The first number determines the representation of rows, and the second number determines how columns are represented. The result is a three-dimensional plot, which opens in a separate window and can be rotated with the help of the left button of the mouse or the touchpad. You can also zoom out and in by holding the right button and scrolling the mouse wheel or using your touchpad. Figure 19.3 displays the same solution in a three-dimensional space. A closer inspection shows that the third dimension slightly draws apart the 'black-and-white' academic and spoken registers.

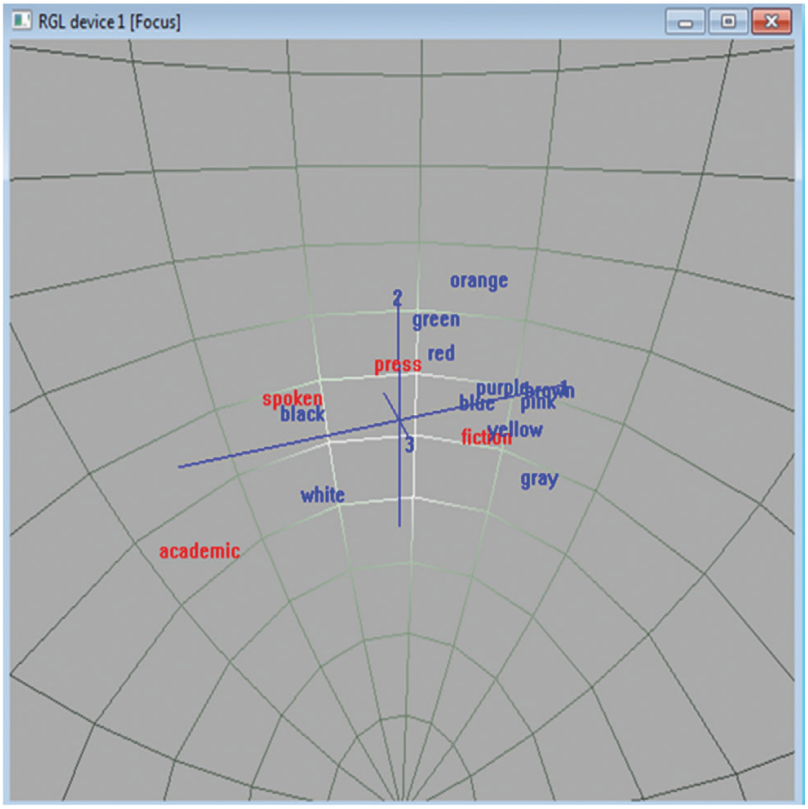

Figure 19.3. A three-dimensional plot of the BCT in four registers

To summarize, SCA has yielded interpretable results. Most secondary BCT, as well as *blue* and *yellow*, cluster together in the same part of the plot where one finds fiction. Note that *blue* and *yellow* are the closest to the secondary terms in Berlin and Kay's (1969) hierarchy.

The location of *green* and *red* is relatively high on Dimension 2, which is also the position of newspaper and magazine texts.

19.2 Visualization of exemplars and prototypes of lexical categories: Multiple Correspondence Analysis of *Stuhl* and *Sessel*

19.2.1 The data and theoretical background

The data and functions mentioned in this case study are available in several add-on packages, which should be installed and loaded:

```
> install.packages(c("FactoMineR", "ca", "rms"))
> library(Rling); library(FactoMineR); library(ca); library(rms)
```

Linguistics, in particular lexicology and semantics, has been influenced by categorization theories in psychology, especially Prototype Theory and, more recently, Exemplar Theory (e.g. Bybee 2001; Taylor 2012). According to Prototype Theory (e.g. Rosch 1975, Rosch & Mervis 1975), categorization of a new item is performed by comparing it with the prototype of an existing category. The prototype is the summary representation of a category. It contains all features of the category instances. Those features are weighted according to their frequency of occurrence in the subject's previous experience. For instance, most instances of the category BIRD have the feature 'can fly', and only some of them have the feature 'can swim'. The feature 'can fly' therefore will be weighted higher than the feature 'can swim'. In this sense, different members of a category can have different degrees of prototypicality, since they possess typical features to a different extent. Thus, a robin, which can fly and cannot swim, is a more prototypical member of the category BIRD than a penguin, who cannot fly but is an excellent swimmer.

Although Prototype Theory has been outperformed by more recent models, such as Exemplar Theory, most categorization theories of the present share a few basic assumptions. One of them is the crucial role of features in establishing the similarity between two exemplars. These features are highly intercorrelated. For instance, a typical bird can fly, has wings, and makes nests in trees. If it did not have wings to keep it in the air, it would not be able to fly. If it were not able to fly, it would make nests in other places.

The case study that illustrates Multiple Correspondence Analysis belongs to the domain of lexical semantics. It focuses on two German lexical categories, *Stuhl* 'chair' and *Sessel* 'armchair'. The categories have been investigated in a classic study by Gipper (1959). In his experiment, German-speaking subjects were asked to name the piece of furniture shown in a picture. Gipper studied the relative frequencies of *Stuhl* and *Sessel* in the subjects' responses. According to his proto-statistical analysis, the two lexical categories

differed regarding the opposition between functionality (*Stuhl*) and comfort (*Sessel*). A theoretically important observation was that the boundaries between the categories were fuzzy, a fact that was against the mainstream structuralist view back then.

The dataset contains 188 instances of these two categories from online stores. The data are available in the dataset `chairs`. The first variable *Shop* represents one of the three online stores; the second one, *WordDE*, shows the exact lexical label of each chair or armchair, and the third variable, *Category*, corresponds to the lexical category (in most cases, the last element of the composite), 'Stuhl' or 'Sessel'. The instances were coded on the basis of online descriptions and pictures for 16 parameters represented by the remaining variables, most of which are self-explanatory.

```
> data(chairs)
> str(chairs)
'data.frame': 188 obs. of 19 variables:
$ Shop          : Factor w/ 3 levels "ikea.de","Moebel-Profi.de",..:
2 1 1 2 1 3 …
$ WordDE        : Factor w/ 44 levels "3-in-1-Sessel",..: 2 17 38
41 23 13 25 15 …
$ Category      : Factor w/ 2 levels "Sessel","Stuhl": 2 2 1 2 2 2
2 1 2 2 …
$ Function      : Factor w/ 5 levels "Eat","NotSpec",..: 1 1 2 1 1
5 2 4 1 1 …
$ Age           : Factor w/ 2 levels "Adult","Children": 1 2 1 1 2
1 1 1 1 1 …
$ Back          : Factor w/ 4 levels "Adjust","High",..: 3 4 4 2 2
2 4 2 4 4 …
$ Soft          : Factor w/ 3 levels "No","Pad","Yes": 1 1 1 3 1 3
1 3 1 1 …
$ Arms          : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 2 1
1 …
$ Upholst       : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 1 2 1
2 …
$ MaterialSeat  : Factor w/ 10 levels "Fabric","Leather",..: 6 10
8 1 6 1 10 2 10…
$ SeatHeight    : Factor w/ 3 levels "Adjust","High",..: 3 2 3 3 2
1 3 3 3 3 …
$ SeatDepth     : Factor w/ 3 levels "Adjust","Deep",..: 3 3 3 3 3
2 3 2 3 3 …
$ Swivel        : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1
1 …
$ Roll          : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1
1 …
```

```
$ Rock          : Factor w/ 2 levels "No","Rock": 1 1 1 1 1 1 1 1
1 1 …
$ AddFunctions  : Factor w/ 3 levels "Bed","No","Table": 2 2 2 2 2
2 2 2 2 2 …
$ Recline       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1
1 …
$ ReclineBack   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 1
1 …
$ SaveSpace     : Factor w/ 3 levels "collapse","No",..: 2 2 3 2 2
2 1 2 2 2 …
```

As one can imagine, many of the variables are intercorrelated. For instance, if a chair can swivel, it can usually roll. One can use the `xtabs()` function to cross-tabulate the variables.

```
> swivelRoll <- xtabs(~ chairs$Swivel + chairs$Roll)
> swivelRoll
     chairs$Roll
chairs$Swivel  No   Yes
No             133  1
Yes            14   40
> chisq.test(swivelRoll)

     Pearson's Chi-squared test with Yates' continuity correction

data: swivelRoll
X-squared = 117.1027, df = 1, p-value < 2.2e-16
```

The χ^2-test confirms our expectations that the variables *Swivel* and *Roll* are associated.

19.2.2 Multiple Correspondence Analysis

To analyse multivariate data with more than two categorical variables, one needs Multiple Correspondence Analysis (MCA). In this case study we will use the `FactoMineR` package, which offers a few useful options for visualization and interpretation of results. This package was also discussed in the previous chapter in the section about PCA.

Recall that the input data for SCA in Section 19.1 was in the form of a contingency table, which cross-tabulated the values of two categorical variables representing registers and Basic Colour Terms. The input data for MCA are normally in the data frame format, where the rows are individual observations and the columns are categorical variables. MCA, in addition to relationships between the values of categorical variables, can also represent individual observations. An individual is located at the same part of the plot as the values of variables it is characterized by. In most cases, one is interested

in the relationships between the variables, i.e. how strongly they are associated and whether they can be interpreted in terms of common dimensions of variation. Plotting the individual members of a category can reveal the category structure, its centre and periphery, clusters of observations, which may correspond to more or less distinct subcategories.

Let us first create a map of 16 variables and the exemplars, using a two-dimensional representation, for reasons that will be addressed below. The first three columns of the data frame will be ignored at this point.

```
> chairs.ca <- MCA(chairs[, -c(1:3)], graph = FALSE)
```

Next, we create a customized version of the MCA map, specifying the size of the labels (`cex = 0.7`), the colours of the variables (`col.var = "black"`) and the colours of individual exemplars (`col.ind = "grey"`). The result is displayed in Figure 19.4.

```
> plot(chairs.ca, cex = 0.7, col.var = "black", col.ind = "grey")
```

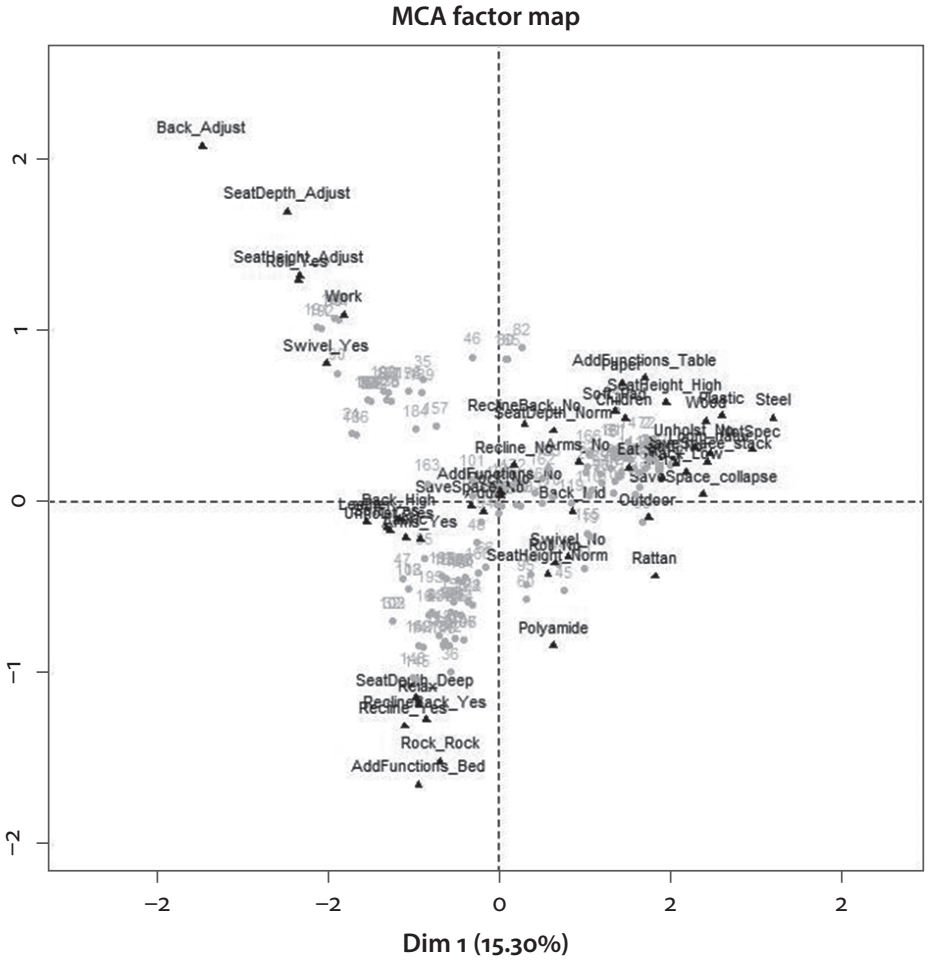

Figure 19.4. Multiple Correspondence Analysis map of *Stuhl* and *Sessel* data with features (black) and exemplars (grey)

The package `FactoMineR` offers many useful tools for interpretation. For example, one can retrieve the contributions of different variables to the first three dimensions by using the function `dimdesc()`, which was introduced in the previous chapter (only the first dimension is shown here):

```
> dimdesc(chairs.ca)
$'Dim 1'
$'Dim 1'$quali
                R2           p.value
Unholst        0.72940952   1.094774e-54
MaterialSeat   0.74518860   3.215782e-48
Function       0.69158437   1.158923e-45
Soft           0.66568141   9.657154e-45
Swivel         0.40875670   5.393205e-23
Roll           0.38348403   2.728416e-21
SeatHeight     0.39565748   5.870717e-21
Back           0.36654364   3.802707e-18
Arms           0.21473392   2.133731e-11
SeatDepth      0.20909906   3.769585e-10
SaveSpace      0.19444992   2.058545e-09
Age            0.06521465   4.047690e-04
ReclineBack    0.06368029   4.764098e-04
Recline        0.04908474   2.246446e-03

$'Dim 1'$category
                   Estimate    p.value
Unholst_No         0.5083627   1.094774e-54
Swivel_No          0.4028109   5.393205e-23
Roll_No            0.4275049   2.728416e-21
NotSpec            0.6439510   2.149389e-16
Arms_No            0.2642196   2.133731e-11
Soft_No            0.3884948   1.476046e-10
SeatDepth_Norm     0.4470056   1.238635e-06
SeatHeight_High    0.5382304   1.615086e-06
Back_Low           0.7371337   1.713799e-06
Eat                0.2341430   2.409594e-05
Plastic            0.3465410   1.930598e-04
Back_Mid           0.3583736   4.033879e-04
Children           0.2361136   4.047690e-04
ReclineBack_No     0.1638759   4.764098e-04
Recline_No         0.1829562   2.246446e-03
Wood               0.2931048   3.379594e-03
SaveSpace_stack    0.2670527   1.904623e-02
Outdoor            0.3004254   4.625690e-02
Back_High          -0.2201617  5.628503e-03
```

```
Recline_Yes          -0.1829562   2.246446e-03
ReclineBack_Yes      -0.1638759   4.764098e-04
Adult                -0.2361136   4.047690e-04
SeatDepth_Adjust     -0.4374452   2.414491e-04
Back_Adjust          -0.8753456   1.980519e-08
SaveSpace_No         -0.5196558   2.724152e-09
Arms_Yes             -0.2642196   2.133731e-11
Soft_Yes             -0.5750871   5.276672e-13
Relax                -0.4635746   1.991433e-15
SeatHeight_Adjust    -0.6819107   7.910288e-18
Fabric               -0.7046046   3.058115e-19
Work                 -0.7149448   4.237029e-21
Roll_Yes             -0.4275049   2.728416e-21
Leather              -0.8350486   6.013338e-23
Swivel_Yes           -0.4028109   5.393205e-23
Unholst_Yes          -0.5083627   1.094774e-54
[output omitted]
```

The first set of numbers in `$'Dim 1'$quali` shows the statistics (R^2 of a linear regression model) that indicates how strongly each variable is associated with each dimension. Only significantly associated variables are displayed. One can see that the variables that are the most closely associated with the first dimension are the presence or absence of upholstery, function (eating, relaxing, work, etc.), softness of the seat, material and ability to move (swivel and roll). The second set of statistics `$'Dim 1'$category` provides information on the directionality of those associations. These are estimates of simple linear regression coefficients. If an estimate is positive (low back, high seat, multifunctional), then the feature will be located in the right-hand part of the plot with positive values of Dimension 1. If it is negative (e.g. made of leather or fabric, with an adjustable back and seat height and designed for work), then the feature will be found on the left. The greater the deviation from zero, the stronger the effect. It seems that the first dimension contrasts highly comfortable office and relaxation (arm)chairs on the left with less comfortable ones on the right.

An inspection of the second dimension reveals that functionality is still a distinctive feature, but this time chairs for relaxation are contrasted with (arm)chairs for work. The third dimension is more difficult to interpret. To summarize, there are three distinct categories of chairs: comfortable chairs for relaxation, comfortable adjustable chairs for work and multifunctional chairs for the household. In addition, in the middle part of the plot one can find comfortable chairs for the dining room, which share some properties of the above-mentioned subcategories.

How are these differences related to the lexical categories under investigation, *Stuhl* and *Sessel*? Similar to PCA, MCA enables one to add supplementary elements. Those are individuals or variables (both categorical and numeric) that do not take part in the creation of the semantic space and do not contribute to the orientation of its dimensions. Those observations or variables are added for the purposes of interpretation. In this case study, *Category*

('Stuhl' or 'Sessel') will be treated as a supplementary variable. For this purpose, we will add an argument `quali.sup = 1`, which will allow for treating the first column in our subset of the data frame (from the third variable on) as a supplementary qualitative (i.e. categorical) variable. Thus, the linguistic form is literally 'mapped' on the meaning represented by the semantic space created by intercorrelated features of the referents.

```
> chairs.ca1 <- MCA(chairs[, -c(1:2)], quali.sup = 1, graph = FALSE)
> plot(chairs.ca1, invis = "ind", col.var = "darkgrey", col.quali.
sup = "black")
```

The models `chairs.ca` and `chairs.ca1` are the same with regard to the coordinates of the points and interpretation of the dimensions. The difference is only in the presence or absence of supplementary variables. The resulting plot, which is shown in Figure 19.5, indicates that *Stuhl* is associated mostly with the features clustered in the upper right quadrant of the map. These features characterize simple practical chairs. However, it is relatively close to the centre because most office chairs for work (see the features in the upper left quadrant) are also categorized as *Stuhl*. The centroid of *Sessel* is located at the bottom, as are the features associated with comfortable chairs for relaxation. Thus, the results suggest that the functional differentiation is crucial. However, it is associated with a number of other referential features.

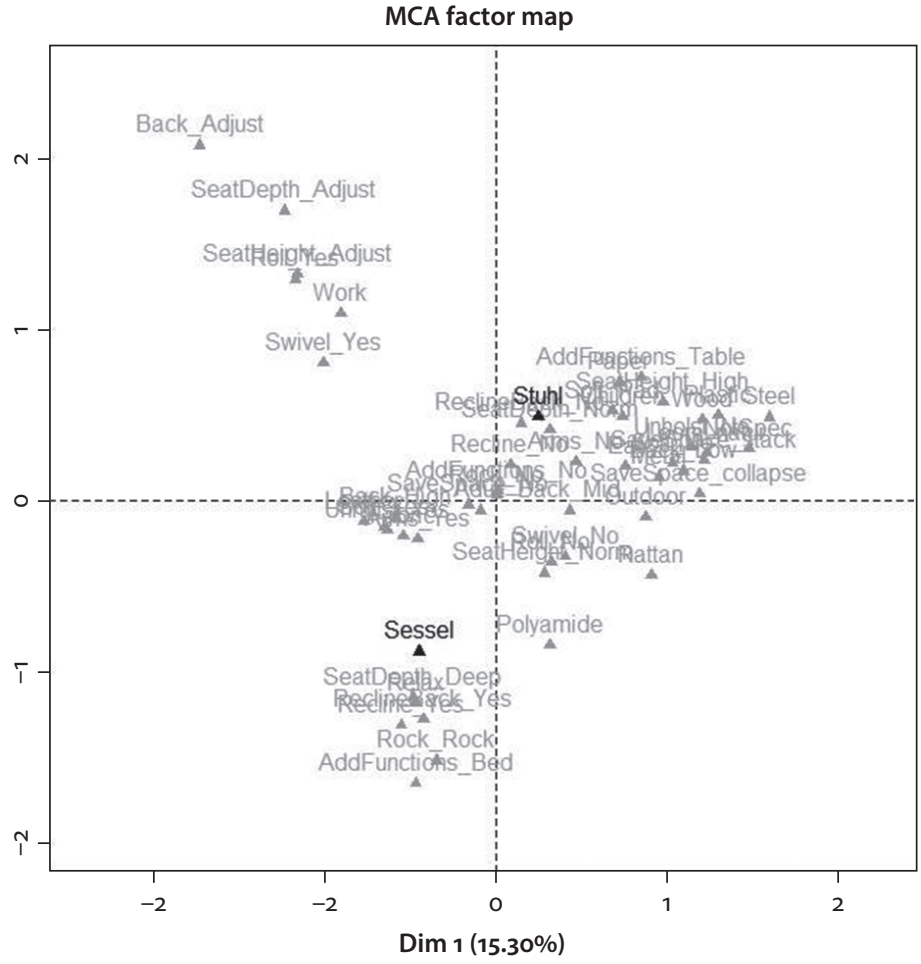

Figure 19.5. Multiple Correspondence Analysis map of *Stuhl* and *Sessel* data with supplementary variables (black) and features (grey)

This implementation of MCA also allows us to construct 95% confidence ellipses around the centroids of *Stuhl* and *Sessel*, which can be regarded as the prototypes of the categories.

```
> plotellipses(chairs.ca1, keepvar = 1, label = "quali")
```

The argument `keepvar = 1` specifies the variable that should be the criterion for classification of the individuals. This variable is *Category*, the first variable in our subset. The final argument `label = "quali"` means that only the labels of the supplementary factor will be plotted.

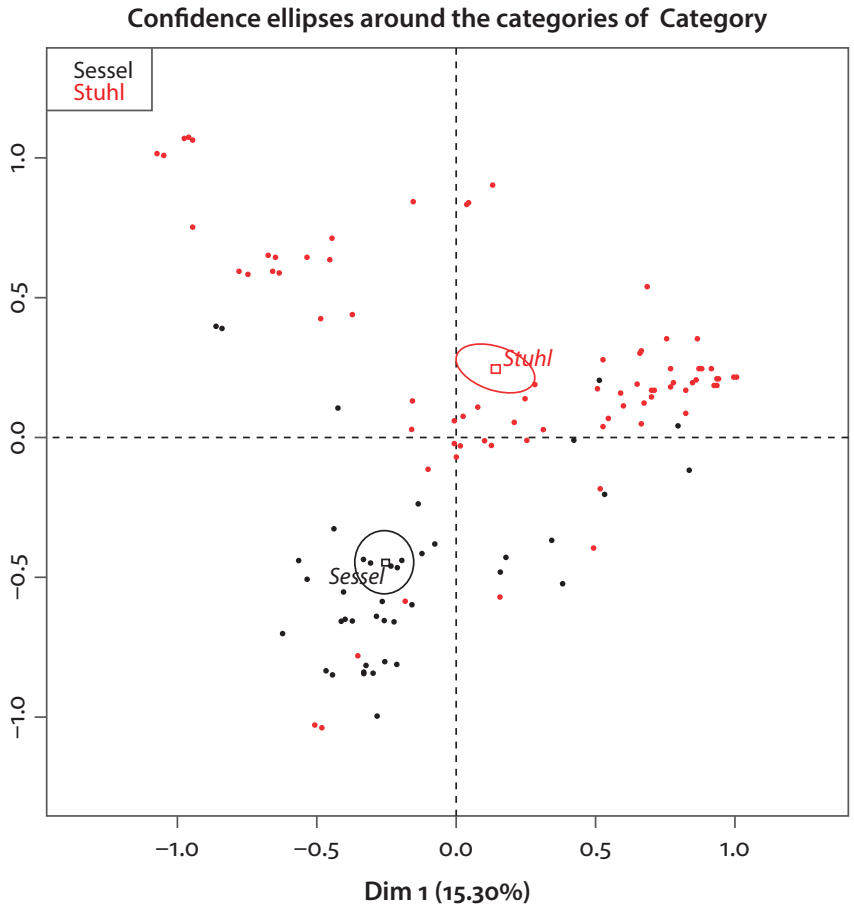

Figure 19.6. Confidence ellipses around the centroids of *Stuhl* and *Sessel*

Figure 19.6 displays the result. Since the confidence ellipses do not overlap, the prototypes can be regarded as distinct. Another option is to create 95% confidence ellipses around all exemplars that represent each category. To choose this option, you should add `means = FALSE`.

```
> plotellipses(chairs.ca1, means = FALSE, keepvar = 1, label =
"quali")
```

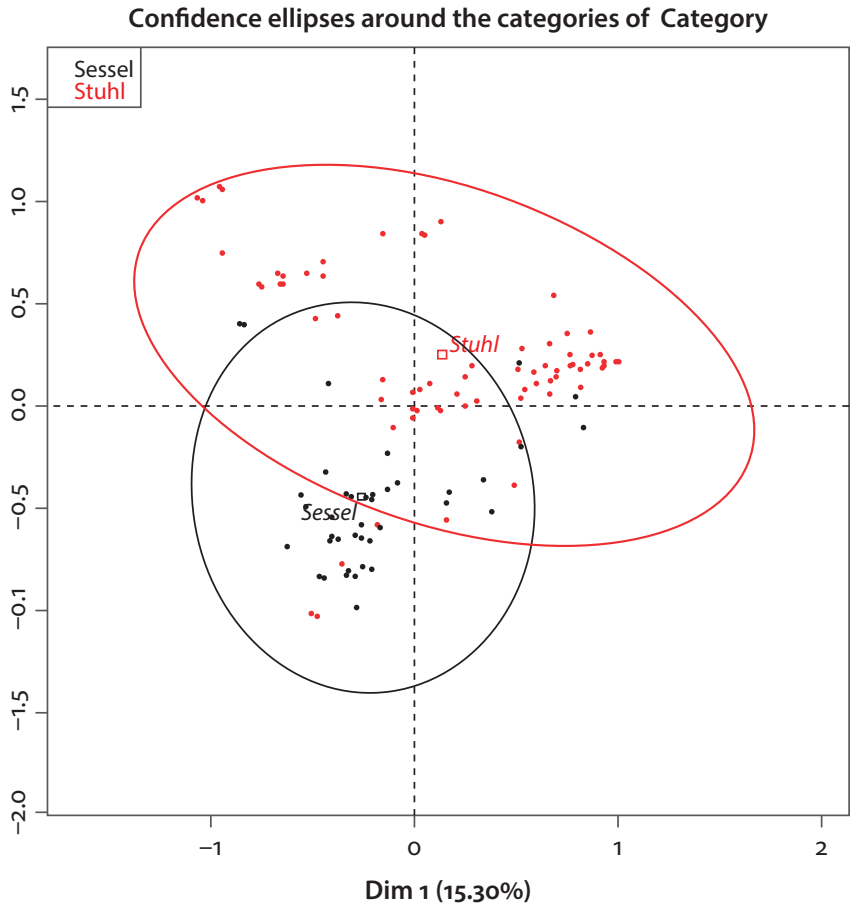

Figure 19.7. Confidence ellipses around the exemplars of *Stuhl* and *Sessel*

Figure 19.7 shows that the categories display a significant overlap, which supports Gipper's observation about the fuzzy boundaries between the categories.

Now it is time to come back to the problem of the number of dimensions. The dimensionality reduction methods discussed in previous chapters usually involve analysis of the proportion of variation explained by the dimensions. In MCA, unlike in Simple Correspondence Analysis, the proportion of explained variance tends to be very modest because the total variance is inflated (see Greenacre 2007: Ch. 19 for a discussion of the problem). The following code yields the eigenvalues, or principal inertias, expressed in absolute values, percentages and cumulative percentages (only the top ten are shown):

```
> chairs.ca$eig
  eigenvalue  percentage  of  variance  cumulative  percentage  of
variance
dim  1  0.3250725720  15.29753280  15.29753
dim  2  0.2576755177  12.12590671  27.42344
dim  3  0.1351901997  6.36189175   33.78533
dim  4  0.1229322264  5.78504595   39.57038
dim  5  0.1089102792  5.12518961   44.69557
```

```
dim   6  0.0961853064   4.52636736    49.22193
dim   7  0.0901939195   4.24441974    53.46635
dim   8  0.0861985147   4.05640069    57.52275
dim   9  0.0816542710   3.84255393    61.36531
dim   10 0.0726465359   3.41866051    64.78397
[output omitted]
```

One can see that the first two dimensions represent only 27.42% of the total variance. That seems to be a very modest result. Greenacre proposes a solution implemented in his `ca` package, namely, adjusted MCA, which estimates explained variation more realistically. Unfortunately, at the moment of writing this solution was not available in `FactoMineR`. One can compare the two-dimensional solution described above with the adjusted version, which is implemented in `mjca()` function in the `ca` package. The adjusted version is default. First, an MCA model of the same data is fit with the help of `mjca()`:

```
> chairs.ca2 <- mjca(chairs[, -c(1:3)])
> summary(chairs.ca2)

Principal inertias (eigenvalues):

dim   value      %     cum%  scree plot
1     0.078443   47.1  47.1  ************************
2     0.043342   26.0  73.2  **************
3     0.006012   3.6   76.8  **
4     0.004155   2.5   79.3  *
5     0.002451   1.5   80.8  *
6     0.001291   0.8   81.5
7     0.000873   0.5   82.1
8     0.000639   0.4   82.4
9     0.000417   0.3   82.7
10    0.000117   0.1   82.8
11    7.6e-050   0.0   82.8
12    1e-05000   0.0   82.8
      ----       ---
Total: 0.166428
[output omitted]
```

The summary suggests that two first dimensions represent 73.2% of inertia (i.e. variance). The horizontal scree plot made of asterisks shows that the subsequent dimensions do not add much to the model. Thus, the two-dimensional solution is correct. But are the non-adjusted and adjusted solutions equivalent? A correlation analysis of the coordinates of features of the first two dimensions (the reader can continue with further dimensions) shows that the solutions are practically identical, differing only in scale and the orientation of the second dimension in `chairs.ca2`, which is turned upside down (hence a negative

correlation). In multivariate exploratory techniques, such 'flipping' is common and should not be a cause of concern (cf. Chapter 17, Section 17.3):

```
> cor(chairs.ca$var$coord[, 1], chairs.ca2$colcoord[, 1])
[1] 1
> cor(chairs.ca$var$coord[, 2], chairs.ca2$colcoord[, 2])
[1] -1
```

Therefore, our initial two-dimensional representation is not perfect, but it is acceptable for a pilot study. Adding more dimensions will not help to improve the fit significantly.

It was mentioned in Chapter 12 that the problem of multicollinearity can be solved by reducing the correlated variables to a smaller set of underlying dimensions. Intercorrelated features of categorical variables are pervasive in linguistic practice. For instance, if a verb depicts a mental state (*think*, *believe*), it is also very likely to have an animate (typically human) first argument (*I believe*), to be followed by a complement clause (*I believe you are right*), and not to be modified by an adverb of speed (**I quickly believe you are right*). Unfortunately, such data are difficult to model with the help of logistic regression, due to data sparseness and multicollinearity. A better option is to do regression on dimensions of MCA. We will demonstrate how one can use MCA dimensions as predictors in a logistic regression model (cf. Chapter 12).

```
> dim1 <- chairs.ca$ind$coord[, 1] #coordinates of individual
exemplars on the horizontal axis
> dim2 <- chairs.ca$ind$coord[, 2] # the same for the vertical axis
> m <- lrm(chairs$Category ~ dim1 + dim2)
> m

Logistic Regression Model
lrm(formula = chairs$Category ~ dim1 + dim2)
     Model Likelihood     Discrimination      Rank Discrim.
              Ratio Test           Indexes           Indexes
Obs          188    LR chi2     118.82   R2     0.643   C      0.921
Sessel       67     d.f.        2        g      2.667   Dxy    0.842
Stuhl        121    Pr(> chi2)  <0.0001  gr     14.394  gamma  0.845
max |deriv|  2e-06                       gp     0.386   tau-a  0.388
                                         Brier  0.094
           Coef    S.E.    Wald Z  Pr(>|Z|)
Intercept  0.9833  0.2448  4.02    <0.0001
dim1       2.1780  0.5319  4.09    <0.0001
dim2       3.9151  0.5377  7.28    <0.0001
```

The results demonstrate that the two dimensions have a high predictive power in the choice between the lexical categories ($C > 0.92$). The coefficients show that the higher the value of

an exemplar with regard to dimension 1 and dimension 2 in the MCA, the greater are the chances of it being categorized as a *Stuhl*. This supports our conclusions that were made after examining the supplementary variables and confidence ellipses.

MCA and Behavioural Profiles

A note should be made about MCA and the Behavioural Profiles approach (see Chapter 15). The methods require identical input data, with exemplars coded for a set of categorical variables. Each method has its advantages and disadvantages. The main advantages of MCA are as follows. First, it allows one to interpret the differences between categories with the help of features shown on the same biplot. Second, an MCA can show both tokens (exemplars) and types (categories) of constructions/words, whereas Behavioural Profile vectors deal only with types. A token-based representation allows one to see the centre and periphery of a category, as well as the overlap between similar categories. An advantage of Behavioural Profiles is that their dendrograms are easier to interpret and validate than the patterns observed in MCA maps. Multidimensional MCA solutions may be especially challenging for interpretation. However, in practice results of Correspondence Analysis tend to be similar to clusters of Behavioural Profiles. See, for example, Newman (2011), who uses both methods to compare verbs of slow movement, such as *trudge*, *plod* and *hobble*.

19.3 Summary

This chapter has discussed Simple and Multiple Correspondence Analysis – techniques developed specially for categorical data. Like PCA and FA, CA can be used to represent data in a few interpretable dimensions. While Simple Correspondence Analysis deals with two cross-tabulated variables, Multiple Correspondence Analysis visualizes the relationships between three and more categorical variables. We have discussed how to create and interpret CA maps, choose the optimal number of dimensions, use supplementary elements, draw 95% confidence ellipses, and other things. The chapter has also shown how one can use the dimensions of MCA as independent variables in regression analysis.

How to report results of Correspondence Analysis

In addition to the plots, one should provide the number of dimensions with the proportions of explained variance (inertia) and features that contribute to the orientation of dimensions. For SCA with two categorical variables, one also reports the χ^2-statistic, the degrees of freedom and the *p*-value (see Chapter 9). For MCA, it is recommended to add the results of a confirmatory logistic regression, if available (see Chapter 12).

More on Correspondence Analysis

Many other kinds of Correspondence Analysis are discussed in detail in Greenacre (2007). For more examples of use of MCA in `FactoMineR`, see Husson et al. (2010).