CHAPTER 18

Multidimensional analysis of register variation

Principal Components Analysis and Factor Analysis

What you will learn from this chapter:

In this chapter you will learn about Principal Components Analysis and Factor Analysis. The aim of these methods is to reduce a large number of correlated quantitative variables to a small set of underlying dimensions. You will learn how to use these methods to perform corpus-based multidimensional analysis of register variation.

18.1 Multidimensional analysis of register variation

Knowledge of the repertoire of registers and linguistic features associated with them is an important part of one's language competence. Registers are language varieties associated with a particular situation of use, e.g. face-to-face conversations, emails, textbooks, fictional novels, lectures and Twitter messages. These situations can be characterized by such parameters as the channel of communication (speech, writing or signing), relationships between the participants (e.g. social status and personal relationships), communicative purpose (transfer of information, persuasion, entertainment, etc.) and settings (e.g. private or public place of communication). On the other hand, registers are also associated with specific linguistic features. For instance, face-to-face conversations contain a higher proportion of first- and second-person pronouns than textbooks, and a lower proportion of nouns and adjectives (e.g. Biber 1988).

In his famous multidimensional analysis of register variation, Douglas Biber (1988) used Factor Analysis to compare diverse texts along a relatively small number of interpretable dimensions of register variation, such as 'Involved versus Informational Production' or 'Narrative versus Non-narrative Concerns'. These dimensions emerge automatically from a large number of variables that describe the proportions of different linguistic features in a set of texts that represent a register. The registers and specific texts can then be mapped onto this register space according to their linguistic features. For instance, face-to-face and telephone conversations will have high scores on the involved pole of the dimension 'Involved versus Informational Production', whereas academic prose will have a high score on the informativity pole.

This type of analysis is usually performed with the help of exploratory Factor Analysis (FA). Since this method is closely related to Principal Components Analysis (PCA), both methods will be discussed in this chapter. Although both of them are used to simplify the data structure and classify variables and objects, FA is more appropriate for detecting theoretically relevant underlying dimensions in the data. However, PCA and FA usually yield similar results, especially if the variables are strongly correlated and the number of variables is large (Field et al. 2012: 760).

18.2 Case study: Register variation in the British National Corpus

18.2.1 The data and research question

You will need several add-on packages for this case study. They should be installed and then loaded.

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

In this section we will carry out a multivariate analysis of register variation. The data are 69 subsections of the BNC coded for 11 variables, which represent normalized frequencies of different parts of speech in each subsection. The subsections represent five broadly defined metaregisters (the spoken part [mostly conversations], fiction, news, academic texts, non-academic prose) plus a 'miscellaneous' category. The data have the following structure:

```
> data(reg_bnc)
> str(reg_bnc)

'data.frame': 69 obs. of 12 variables:
$ Reg        : Factor w/ 6 levels "Acad","Fiction",..: 6 6 6 6 6 6
6 6 6 6 …
$ Ncomm      : num 0.17 0.205 0.206 0.136 0.133 …
$ Nprop      : num 0.02697 0.02498 0.0468 0.0112 0.00985 …
$ Vpres      : num 0.0355 0.0391 0.0366 0.0485 0.0452 …
$ Vpast      : num 0.0219 0.0298 0.0236 0.0189 0.0198 …
$ P1         : num 0.0347 0.0208 0.018 0.0276 0.0455 …
$ P2         : num 0.01832 0.01137 0.00775 0.03749 0.03703 …
$ Adj        : num 0.0536 0.0585 0.0596 0.0407 0.0446 …
$ ConjCoord  : num 0.0395 0.034 0.0335 0.0339 0.0384 …
$ ConjSub    : num 0.031 0.0276 0.0232 0.0315 0.0283 …
$ Interject  : num 0.00997 0.00414 0.00226 0.02173 0.04298 …
$ Num        : num 0.0206 0.0192 0.0277 0.0414 0.0164 …
```

All columns except the first one are numeric vectors that represent the proportions of various word classes in 69 BNC subsections. The latter are provided as the row names:

```
> rownames(reg_bnc)
[1] "S_brdcst_disc" "S_brdcst_doc" "S_brdcst_news" "S_classroom"
[5] "S_consult"     "S_conv"       "S_courtroom"   "S_demonstratn"
[9] "S_interv_oral" "S_interview"  "S_lect_arts"   "S_lect_com"
[output omitted]
```

The main question of this study is as follows. Can we identify interpretable dimensions of register variation on the basis of the data and what are the relationships between the metaregisters, as well as more specific subsections, with regard to these dimensions? To answer this question, we will use PCA and FA. We will begin with PCA because it is conceptually less complex.

18.2.2 Principal Components Analysis

Before beginning the analyses, it is useful to check whether the data are actually appropriate for PCA and FA. There are two important conditions: on the one hand, the variables should be intercorrelated (otherwise, we cannot reduce the data to a smaller number of underlying components); on the other hand, the correlations should not be too high. Very high correlations are not a problem for PCA, but they can cause inaccurate estimates in FA, similar to multicollinearity in multiple regression (Field et al. 2012: 770–771). As a very approximate rule of thumb, the absolute values of correlations should not be lower than 0.3 and above 0.9. To examine the correlations between the variables, one can use `cor()`:

```
> round(cor(reg_bnc[, -1]), 2)
           Ncomm  Nprop  Vpres  Vpast  P1     P2     Adj
Ncomm      1.00   0.23   -0.41  -0.21  -0.83  -0.75  0.86
Nprop      0.23   1.00   -0.34  0.36   -0.37  -0.50  0.13
Vpres      -0.41  -0.34  1.00   -0.46  0.42   0.50   -0.35
Vpast      -0.21  0.36   -0.46  1.00   0.03   -0.11  -0.16
P1         -0.83  -0.37  0.42   0.03   1.00   0.80   -0.79
P2         -0.75  -0.50  0.50   -0.11  0.80   1.00   -0.70
Adj        0.86   0.13   -0.35  -0.16  -0.79  -0.70  1.00
ConjCoord  -0.13  -0.45  0.21   0.07   0.23   0.31   0.04
ConjSub    -0.52  -0.68  0.48   -0.22  0.57   0.57   -0.39
Interject  -0.67  -0.39  0.41   0.02   0.70   0.79   -0.62
Num        0.21   0.28   -0.28  -0.13  -0.25  -0.16  0.03
           ConjCoord  ConjSub  Interject  Num
Ncomm      -0.13      -0.52    -0.67      0.21
Nprop      -0.45      -0.68    -0.39      0.28
Vpres      0.21       0.48     0.41       -0.28
Vpast      0.07       -0.22    0.02       -0.13
P1         0.23       0.57     0.70       -0.25
```

```
P2          0.31       0.57      0.79       -0.16
Adj         0.04       -0.39     -0.62      0.03
ConjCoord   1.00       0.26      0.18       -0.41
ConjSub     0.26       1.00      0.36       -0.28
Interject   0.18       0.36      1.00       -0.09
Num         -0.41      -0.28     -0.09      1.00
```

The variable *Num* has many correlations with the absolute value slightly under 0.3. One may consider removing such variables. As for possible multicollinearity, this should not be a concern, since there are no highly correlated variables.

One can also use the Bartlett test for this kind of preliminary diagnostics. The null hypothesis of this test is that the variables are not correlated. In that case, it would not make sense to run further analysis. The test is available as `cortest.bartlett()` in the package `psych`.

```
> cortest.bartlett(reg_bnc[, -1])
R was not square, finding R from data
$chisq
[1] 536.3401

$p.value
[1] 4.109611e-80

$df
[1] 55
```

Since the *p*-value is well below the significance level, we can reject the null hypothesis of zero correlation between the variables and continue with the analyses.

We will begin with PCA, using `PCA()` from the `FactoMineR` package:

```
> reg.pca <- PCA(reg_bnc, quali.sup = 1, graph = FALSE)
```

The argument `quali.sup = 1` tells R that the first variable, *Reg*, should be regarded as a qualitative supplementary variable. In contrast to active elements, supplementary elements do not contribute to the construction of principal components. They are added to the analysis for the purpose of interpretation or illustration. There are two types of supplementary variables, qualitative and quantitative. Next, the argument `graph = FALSE` suppresses the immediate creation of graphical output. Note that PCA is usually performed on standardized scores (see Chapter 3) rather than original ones because it is sensitive to scaling differences. This is the default option, which should not be changed.

The first question is how many dimensions are needed to account for register variation. The output of `reg.pca$eig` shows eigenvalues, proportions of variance explained by each component and cumulative explained variance. The eigenvalue is an important concept in multivariate analysis. As was mentioned in the previous chapter, an eigenvalue shows how much of the total variance is explained by each component. The higher the

correlations between a component and the variables, the greater the component's eigenvalue. Note that the eigenvalue of every additional component is smaller than the previous one. The first component always explains the greatest portion of variance. However, the cumulative percentage of explained variance always increases, until it reaches 100%.

```
> head(reg.pca$eig)
  eigenvalue  percentage  of  variance  cumulative  percentage  of
  variance
comp  1  5.0682936  46.075396  46.07540
comp  2  1.8722103  17.020094  63.09549
comp  3  1.3758435  12.507669  75.60316
comp  4  0.7900757  7.182506   82.78566
comp  5  0.6451271  5.864791   88.65046
comp  6  0.4217144  3.833768   92.48422
```

There are different rules of thumb with regard to the optimal number of components: some statisticians believe that one should retain only those components whose eigenvalues are greater than 1 (the Kaiser criterion); others are less strict, and use 0.7 as a cut-off point (see a discussion in Field et al. 2012: 762–764). One can also inspect the values visually using a scree plot, as has been done in the previous chapters. Consider Figure 18.1. The *x*-axis of the bar plot represents the number of components from 1 to 11 (the number of rows in `reg.pca$eig`), which were created by the algorithm. The *y*-axis shows the percentage of variance explained by each dimension. These values can be found in the second column in `reg.pca$eig`. The R code is as follows:

```
> barplot(reg.pca$eig[,2], names = 1:nrow(reg.pca$eig), xlab =
"components", ylab = "Percentage of explained variance")
```

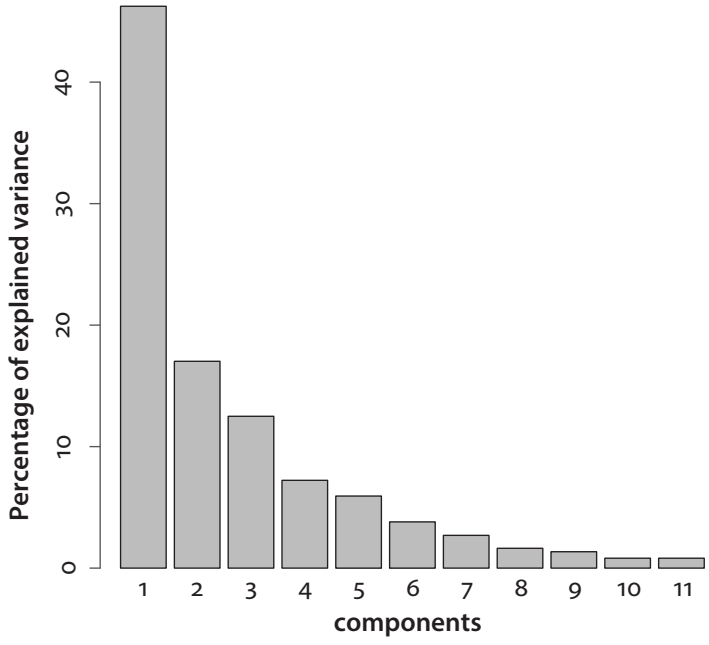

Figure 18.1. Contributions of components of PCA to explaining variance

Figure 18.1 shows no substantial increase in explanatory power after three dimensions, which also explain together 75.6% of variance. This is also the number of components according to the Kaiser criterion, since only the first three components have eigenvalues greater than 1.

Now we can begin interpreting the components, or dimensions. To visualize the variable space, one can use the following command:

```
> plot(reg.pca, choix = "var", cex = 0.8)
```

The argument `choix = "var"` is used to represent the variables. The result is displayed in Figure 18.2. By default, the algorithm creates a plot with individuals (in our case, the BNC subsections). Unless specified otherwise, the plot displays the first two components as two axes. The variables are represented as vectors pointing away from the origin. The angles between the vectors and the axes indicate how strongly the variables are correlated with the dimensions. The smaller the angle, the stronger the correlation. If two vectors point to almost the same direction, this means that the corresponding variables are highly correlated and therefore may represent the same underlying theoretical construct. The length of the vectors reflects how much variation in the variable is captured by this low-dimensional display, with the maximum length of 1 (limited by the circle). In other words, the length represents the quality of the representation of a variable on the plane.

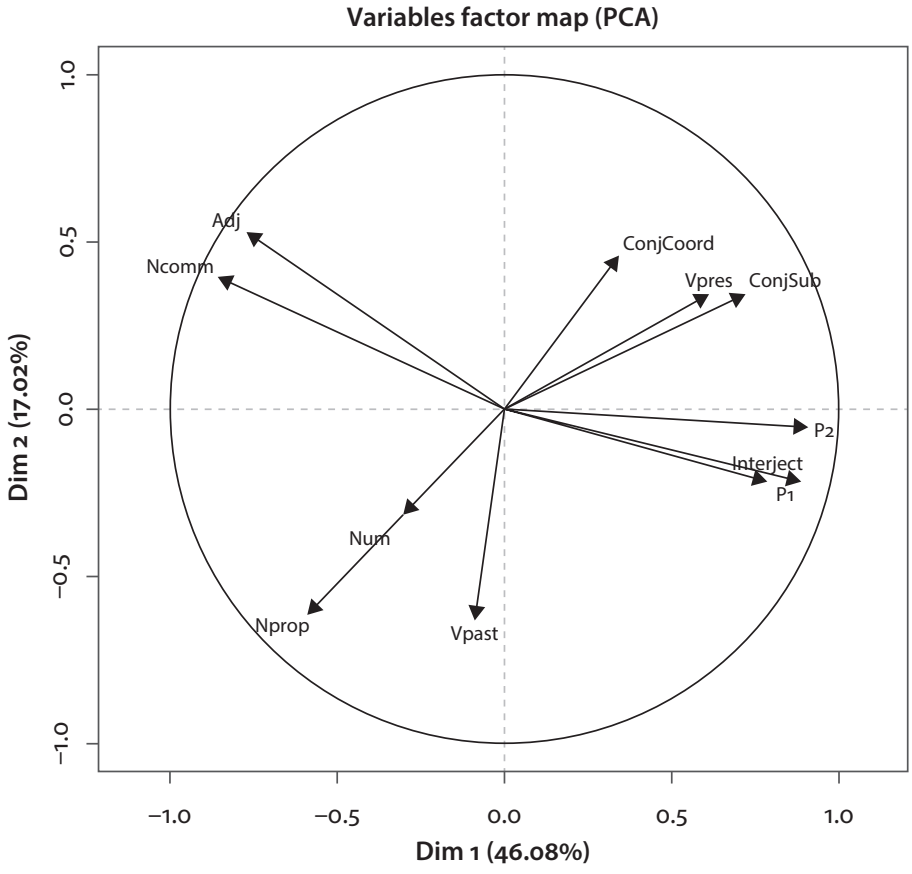

Figure 18.2. The first and second components of Principal Components Analysis with variables

The orientation of variables suggests that the first (horizontal) component relates to the well-known opposition between involved and informational communication. The variables which are the most strongly correlated with the dimension are the ones that represent the frequencies of the 1st and 2nd person pronouns and interjections. These are linguistic features associated with involved, interactive communication. They are opposed to common nouns and adjectives on the left – features that indicate high informational density of communication.

One can also examine the correlation coefficients by using `dimdesc()`. For the first dimension, the output is as follows:

```
> dimdesc(reg.pca)
$Dim.1
$Dim.1$quanti
           correlation  p.value
P2         0.9117524    0.000000e+00
P1         0.8958585    0.000000e+00
Interject  0.7913207    6.661338e-16
ConjSub    0.7268571    1.540101e-12
Vpres      0.6203029    1.311435e-08
ConjCoord  0.3461531    3.574209e-03
Num        -0.3236023   6.681150e-03
Nprop      -0.5825157   1.513793e-07
Adj        -0.7699620   1.056008e-14
Ncomm      -0.8551296   8.636119e-21

$Dim.1$quali
     R2          p.value
Reg  0.7783745   2.391373e-19

$Dim.1$category
            Estimate   p.value
Spok        3.100121   6.960014e-16
Fiction     1.259459   2.530881e-02
Acad        -1.256963  2.978404e-03
NonacProse  -1.272090  2.671066e-03
News        -1.427441  4.102683e-06
[output omitted]
```

The numbers support our previous observations based on the graphical representation: the top positively correlated features are the 1st and 2nd person pronouns and interjections. Their correlations are very strong, around 0.9. The strongest negative correlations are observed for common nouns and adjectives (–0.86 and –0.77, respectively). By default, the function returns only those estimates that are significant at the level of 0.05.

In addition, the function returns the estimates of regression coefficients for qualitative supplementary variables (in this example, the metaregisters). The response variable is

the coordinates of the megaregisters on the dimension. The largest positive estimate (3.1) is observed for the spoken data. This means that it is the most strongly associated with involvement. The spoken data are followed by fiction. The values of the non-fiction written registers are negative, which reflects the informative orientation of the latter.

The second component has relatively high positive values for adjectives and coordinate conjunctions, and negative values for past forms of verbs and proper nouns. One can interpret this dimension tentatively as description vs. reporting of past events. The academic texts are significantly correlated with the positive values. The register that is significantly associated with the negative values on the second dimension is the news. As for the third dimension, it seems to correspond to the distinction between narrative and non-narrative texts. The strongly associated features are past tense verbs and coordinating conjunctions (positive correlation) and numerals (negative correlation). It distinguishes fiction from all other registers. One can visualize the second and third dimensions in a plot with the following command (see Figure 18.3):

```
> plot(reg.pca, axes = c(2,3), choix = "var", cex = 0.8)
```

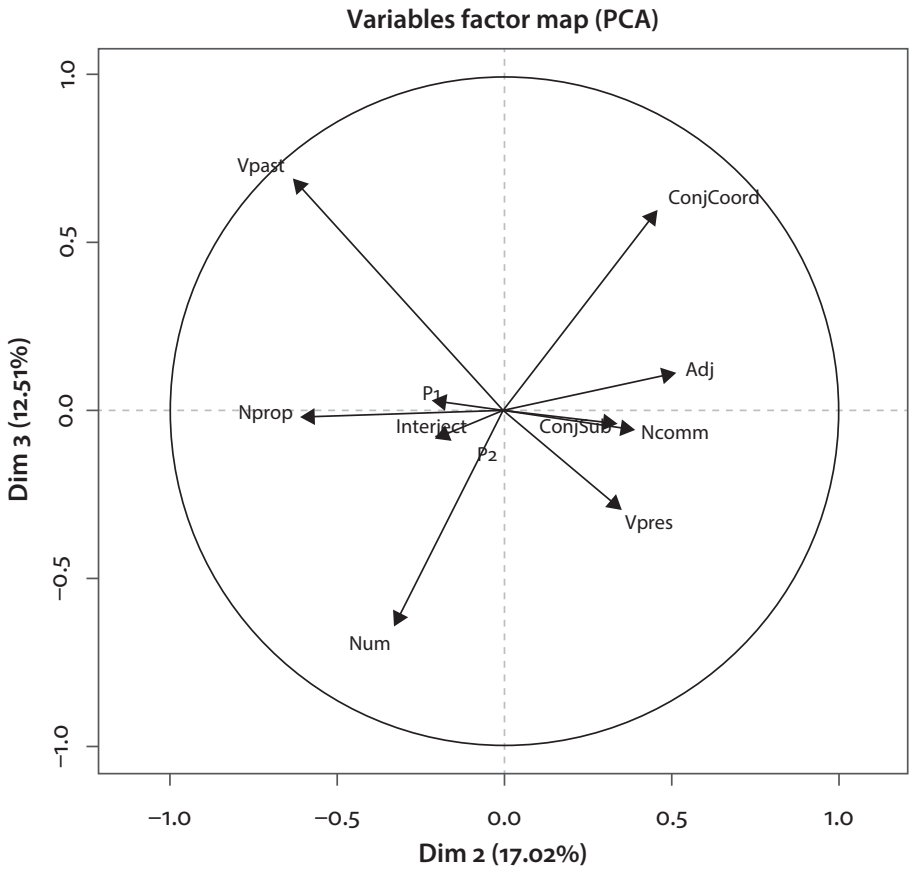

Figure 18.3. The second and third components of the Principal Components Analysis with variables

Finally, we will plot the individual subsections onto this space. Their labels will be in grey, size 0.8. The centroids of the five metaregisters are plotted, as well. Their positions can

be interpreted as the prototypes of the corresponding metaregisters. To plot components 1 and 2, the following code can be used:

```
> plot(reg.pca, cex = 0.8, col.ind = "grey", col.quali = "black")
```

Plotting components 2 and 3 is done as follows:

```
> plot(reg.pca, axes = c(2, 3), cex = 0.8, col.ind = "grey", col.
quali = "black")
```

The results are shown in Figures 18.4 and 18.5. They show that the BNC subsections that belong to the same metaregister tend to cluster together.

The results seem to support one's intuitive ideas about registers. For example, spoken discourse is highly involved (component 1), neutral with regard to description/past event reporting (component 2) and slightly more non-narrative than narrative (component 3).

It is also possible to plot confidence ellipses around the centroids to estimate the amount of overlap of the prototypes of the registers (dimensions 1 and 2):

```
> plotellipses(reg.pca, label = "quali")
```

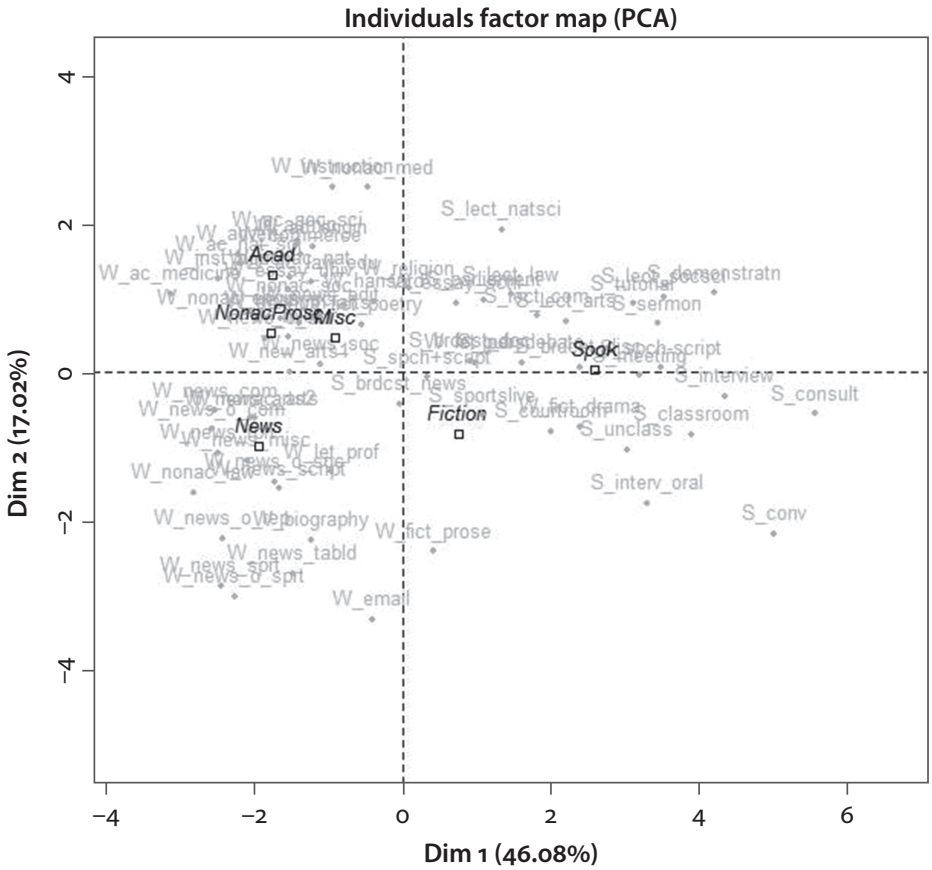

Figure 18.4. Orientation of 69 BNC subsections with regard to Principal Components 1 and 2. Grey text labels: subsections. Black text labels: register centroids

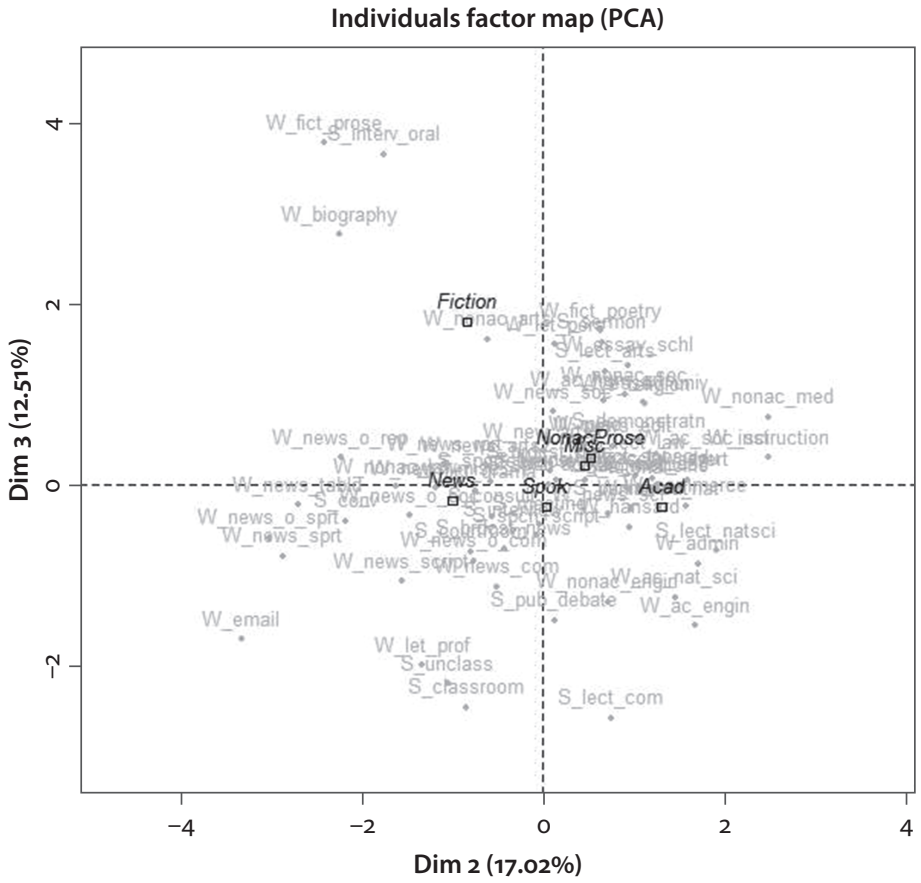

Figure 18.5. Orientation of 69 BNC subsections with regard to Principal Components 2 and 3. Grey text labels: subsections. Black text labels: register centroids

The argument `label = "quali"` tells R to plot only the labels of supplementary elements. Figure 18.6 shows the 95% confidence ellipses of the centroids. Note that the size of confidence ellipses depends on the number of points. The fiction ellipsis is very large, since this category is represented only by three BNC subsections. The large ellipsis of fiction overlaps with the ellipsis of the spoken data, and the category 'miscellaneous' overlaps with most other registers. The reader is encouraged to explore the second and third dimensions by adding `axis = c(2, 3)`. One can also compare the confidence ellipses around all individual subcorpora that represent a register, rather than around the register centroids by adding `mean = FALSE`.

To summarize, even though PCA is not a technique *par excellence* for discovering underlying constructs, we have found three more or less interpretable dimensions based on eleven linguistic variables. These dimensions capture important differences between the registers and text types. The `FactoMineR` package contains many more useful functions, and the reader is encouraged to explore them (see also Husson et al. 2010).

However, it is not entirely clear how dimension 2 is related to dimension 3. Dimension 2 opposes adjectives, coordinate conjunction and common nouns to past tense verbs and proper nouns, whereas Dimension 3 gives positive scores to past tense verbs and

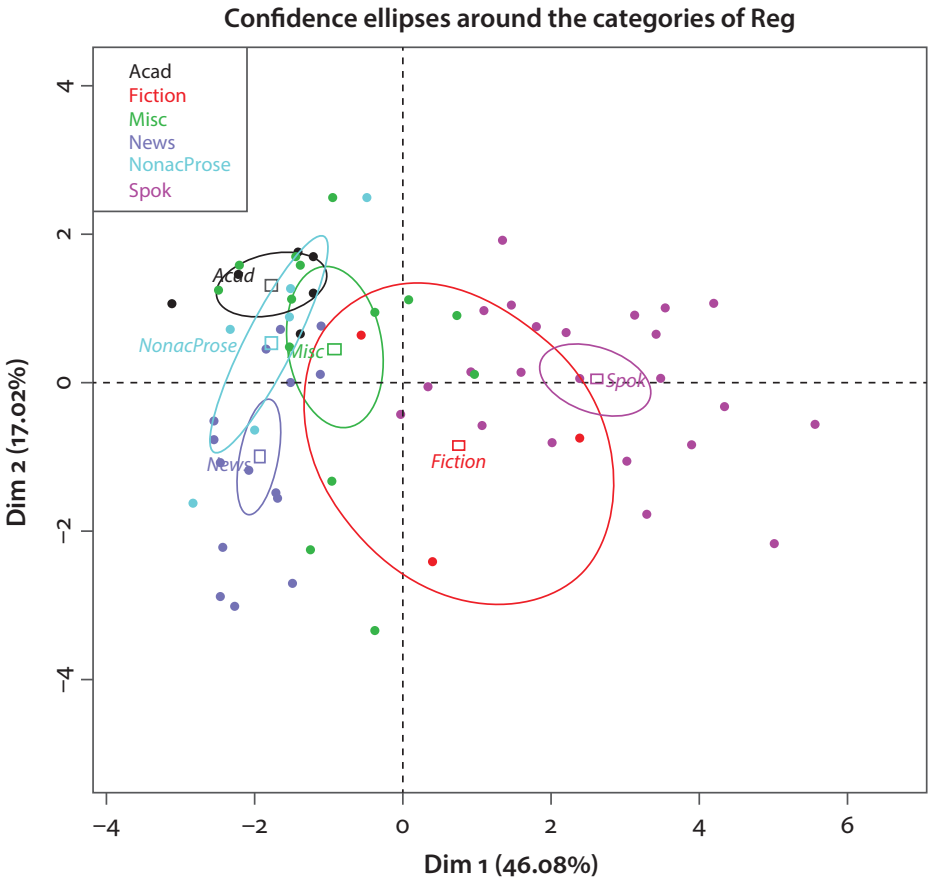

Figure 18.6. Confidence ellipses around the centroids of the registers, Principle Components 1 and 2

coordinate conjunctions, and negative scores to numerals. To achieve greater interpretability, one can try a different method, Factor Analysis, which is a more appropriate tool for finding theoretically interpretable dimensions of variation.

18.2.3 Factor Analysis

The aim of both PCA and FA is to simplify the structure of a set of variables. Yet, these methods have important differences. The main purpose of PCA is to find as few orthogonal (uncorrelated) components as possible while maximizing the total explained variance. It is used mainly to reduce dimensionality. In contrast, FA is more widely used for exploring theoretical constructs, or latent variables, which are called factors. There is also a major technical difference between the methods. Unlike PCA, FA 'rotates' the factors, trying to increase the load of variables on several common factors.

The main function for FA is `factanal()` in the basic R distribution. To perform FA, one has to specify the desired number of factors. We will use the optimal number of components in PCA, that is, three factors.

```
> reg.fa <- factanal(reg_bnc[, -1], factors = 3)
> reg.fa
```

```
Call:
factanal(x = reg_bnc[, -1], factors = 3)

Uniquenesses:
   Ncomm   Nprop   Vpres   Vpast   P1      P2      Adj     ConjCoord
   0.120   0.335   0.510   0.005   0.175   0.192   0.102   0.496
   ConjSub   Interject   Num
   0.438     0.416       0.726

Loadings:
             Factor1  Factor2  Factor3
Ncomm        -0.927            -0.125
Nprop        -0.214   -0.458   -0.640
Vpres         0.417    0.539    0.159
Vpast         0.138   -0.983
P1            0.868    0.118    0.240
P2            0.796    0.259    0.327
Adj          -0.940            0.107
ConjCoord                      0.709
ConjSub       0.480    0.336    0.467
Interject     0.716    0.101    0.248
Num          -0.109            -0.508

                 Factor1  Factor2  Factor3
SS loadings      4.127    1.682    1.676
Proportion Var   0.375    0.153    0.152
Cumulative Var   0.375    0.528    0.680

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 65.18 on 25 degrees of freedom.
The p-value is 1.95e-05
```

An important concept in FA is **factor loadings**. The stronger a variable loads onto a factor, the more strongly this variable defines the factor. Factor loadings are analogous to correlation coefficients between the variables and factors, although their numerical values can be sometimes greater than 1 or less than –1. As a rule of thumb, loadings with absolute (positive or negative) values greater than 0.3 are considered to be important. One can see that the first factor is very similar to the first Principal Component from the previous subsection, with strong positive correlations with the proportions of 1st and 2nd person pronouns and interjections, and strong negative correlations with the proportions of common nouns and adjectives. Thus, Factor 1 represents the distinction between involved and informational communication. This is also the factor that accounted for most variation in Biber's (1988) analysis. Factor 2, which distinguishes between past tense verbs and present tense verbs, looks similar to Biber's (1988) narrative vs. non-narrative dimension. Finally, Factor 3 contrasts coordinate (followed at some distance by subordinate) conjunctions, on the one hand, and proper nouns and numerals, on the other hand. This dimension looks

intriguing. Let us have a look at the subcomponents and their scores. To make the algorithm return scores for the subcomponents, one can add `scores = "Bartlett"`.

```
> reg.fa <- factanal(reg_bnc[, -1], factors = 3, scores = "Bartlett")
> plot(reg.fa$scores[, 2:3], type = "n")
> text(reg.fa$scores[, 2:3], rownames(reg_bnc), cex = 0.7)
```

The resulting plot is shown in Figure 18.7.

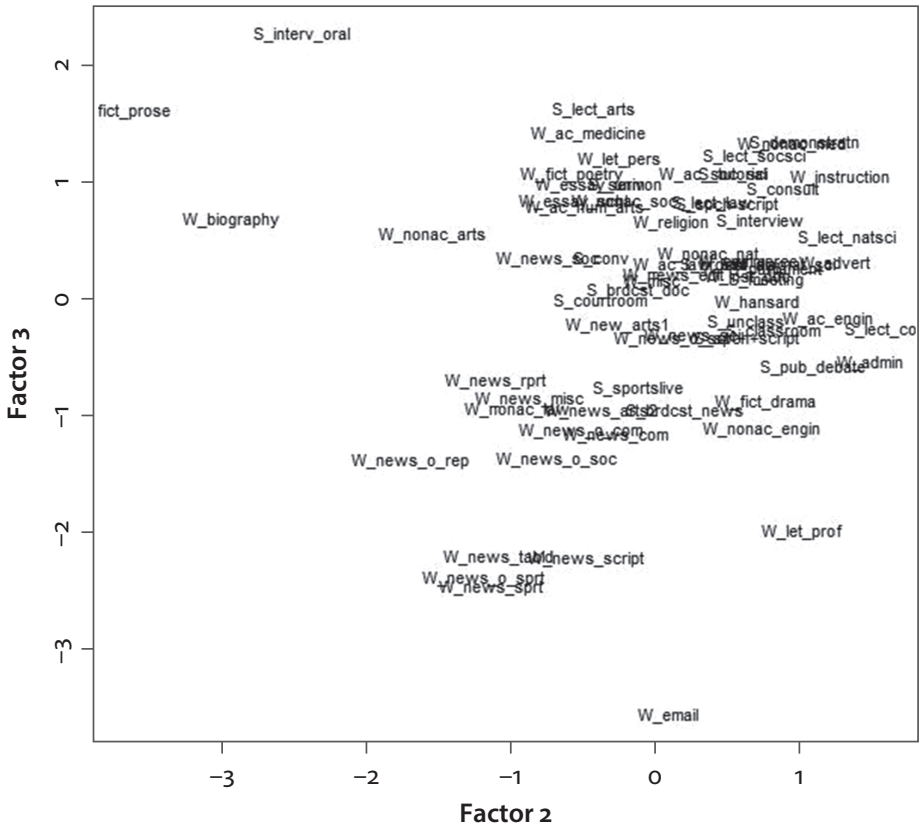

Figure 18.7. Scores of BNC subcorpora on Factor 2 and 3

It seems that the third factor relates to more or less elaborate development of ideas (fiction, interviews, lectures on arts) vs. concise informationally dense communication (news, emails).

When performing FA, one also has to address the question whether three factors are enough. This information can be obtained from the *p*-value in the bottom line of `reg.fa`. A *p*-value smaller than 0.05 indicates that the number of factors is insufficient. Unfortunately, there is much variation in the data that is left unexplained. Increasing the number of factors to six saturates the model ($p = 0.0738$), but does not help to arrive at a parsimonious theoretically interesting solution.

Finally, it is necessary to mention that there are two popular kinds of rotation, called Varimax and Promax. In the case of Varimax (a type of orthogonal rotation), factors are considered uncorrelated, whereas Promax (a type of oblique rotation) allows for some degree of correlation between factors. It is believed that behavioural data are more 'messy'

and factors tend to be correlated. If we expect to find really clear-cut, unique factors, it is better to use Varimax, which is the default option. If one expects the resulting factors to be closely related, one can try Promax. To explore the latter option, one can use the following code:

```
> reg.fa <- factanal(reg_bnc[, -1], factors = 3, rotation = "promax")
> reg.fa$loadings

Loadings:
            Factor1  Factor2  Factor3
Ncomm       -1.005           -0.195
Nprop                -0.718   0.258
Vpres       0.331            -0.472
Vpast       0.259    0.113   1.066
P1          0.869
P2          0.735    0.205
Adj         -1.106   0.350    -0.102
ConjCoord   -0.220   0.853    0.216
ConjSub     0.314    0.448    -0.159
Interject   0.697    0.132
Num                  -0.596   -0.238

                 Factor1  Factor2  Factor3
SS loadings      4.347    2.012    1.615
Proportion Var   0.395    0.183    0.147
Cumulative Var   0.395    0.578    0.725
```

One can see that Factors 2 and 3 have swapped their order in comparison with the previous version, but otherwise the results are quite similar. Usually, Promax yields a better fit. Note that cumulative explained variance is greater than it was in the previous solution.

How reliable is your questionnaire?

Factor Analysis is frequently used on questionnaire data. But how reliable is the questionnaire? Does it reflect the underlying construct well? A popular measure of reliability is the Cronbach α ('alpha').

One could think of such questions as 'Do you like learning new languages?', 'Do you read books in foreign languages?' and 'Do you try to speak the local language when you stay in another country?' as measures of one's interest in foreign languages. To measure

how related these questions are and how well they represent the underlying construct, one can use the `alpha()` function in the package `psych`:

```
> alpha(cbind(Question1, Question2, Question3)) # do not run
```

where the variables `Question1`, `Question2` and `Question3` are vectors with the subjects' responses on a scale. If some questions imply a reversed scale, one can specify that with the `keys` argument (see the help page). The function returns different versions of the coefficient, but the most traditional one is `raw_alpha`. The closer its value to one, the better, although extremely high values may suggest that the questions are tautological.

18.3 Summary

This chapter has introduced the basics of Principal Components Analysis and Factor Analysis. A multidimensional analysis of register variation in the BNC served as an illustration. With the help of these closely related methods we have managed to find interpretable dimensions of register variation. The chapter has also demonstrated how relationships between different text types/registers and linguistic variables can be explored with the help of biplots (i.e. plots that represent both rows and columns of a dataset in one common space). Of course, the results of this case study cannot be regarded as conclusive. The 'real' multidimensional analysis of registers requires many more linguistic variables than were considered in our case study. For example, Biber (1988) used almost 70 features.

Multidimensional analysis of register variation is not the only possible application of PCA and FA in linguistics. For example, one can use loadings of components or factors as input in regression analysis to solve the problem of multicollinearity or simplify the model.

How to report results of PCA and FA

Reporting PCA and FA is usually quite verbose. One should mention the sample size, the number of variables and the procedure, namely, how you made the decision about the number of components/factors, as well as the rotation method and the *p*-value. Crucially, one should include a table with factor loadings per each variable. Of course,

(*Continued*)

all relevant biplots should be provided, as well, if the purpose is also to obtain a classification of observations.

More on PCA and FA

A more traditional function for performing PCA is `princomp()` in the base package. To learn about other variants of rotation in FA (such as Quartimax, Oblimin, BentlerT, etc.) and other options, see the help page of `fa()` in the package `psych`. For an accessible explanation of the theory behind PCA and FA, see Field et al. (2012: Ch. 17). This chapter dealt only with exploratory FA, leaving out confirmatory FA. The latter is more complex and is typically used to test how well a hypothesized structure fits a set of data. For an introduction to confirmatory FA with R, see Everitt & Hothorn (2011: Ch. 7).