PCA Table of Content

PRINCIPAL COMPONENT ANALYSIS: THE BASIC

Hervé Abdi, Carlos Gomez, Vincent Guillemot, Soudeh A. Khoubrouy & Ju-Chi Yu
2024-07-23

Principal component analysis (PCA for short) is, by far, the oldest multivariate method: The mathematical formulation can be found as early as in 1829 (by the French mathematician Cauchy, see Abdi and Williams, 2010), whereas the geometrical formulation can be found in a paper published in 1901 by Karl Pearson (of χ^2 fame). But the modern approach is exposed in 1933 by Hotteling and in 1936 by Eckart and Young. Most of the current multivariate methods are variations over the theme of PCA which remains (by far) the most popular multivariate method.

PCA is an exploratory method that analyzes a data table for which some observations (that are in general the rows of the data table) are described by quantitative variables (that are in general the columns of the data table). PCA extracts, from this data table, the relevant information and represents it with maps (sometimes called factorial or factor maps) that show on one hand the structure of the observations and on the other hand the structure of the variables.

In sensory evaluation, PCA is mostly used to analyze data for which products (the observations) are described by some sensory attributes (the variables)—A pattern that corresponds, for example, to experimental procedures such as Sensory Profile or Rate All That Apply (RATA) or Polarized Sensory Positioning.

1. The data for PCA

Figure 1.1 shows an example of a data table suitable for a PCA. Here the rows of the data table are beers and the columns are sensory characteristics (e.g., *Bitter, Acidic*). At the intersection of a row (representing a beer) and a column (representing a sensory characteristic) we find the intensity of this sensory characteristic (the column) for this beer (the row).

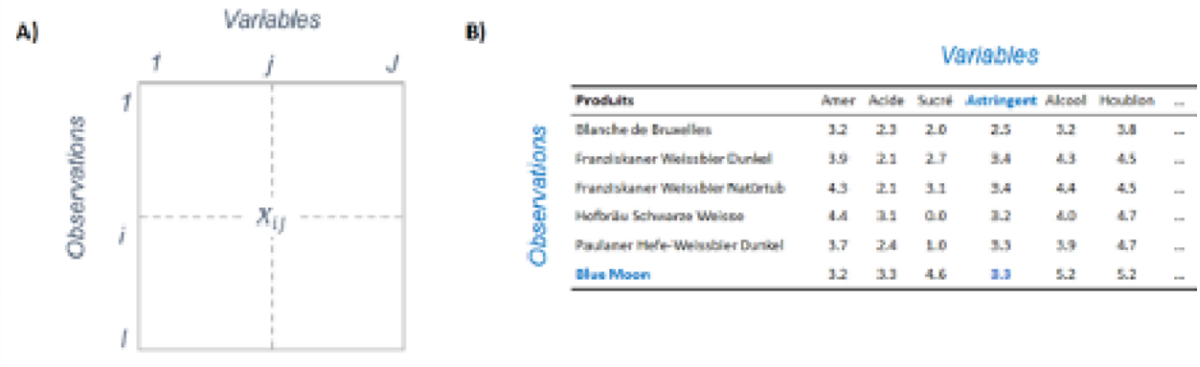

Figure 1.1 An Example of Data Suitable for PCA

2. The goals of PCA

In PCA we have different questions for the observations and the variables of the table. For the observations (i.e., here the products) we want to identify the observations that are similar to each other and those that differ from each other. This way, we want to obtain groups of observations that are described in a similar way by the variables. To do so we need a measure of similarity or its inverse, a measure of dissimilarity. In PCA, the dissimilarity between two observations is evaluated by the squared Euclidean distance between these two observations and the *similarity* is just the inverse of the *dissimilarity*.

For the variables (here the descriptors), the goal is to evaluate the configuration of their interactions—as measured by their covariance or their correlation. This way we want to identify the variables that provide similar information (and could be considered redundant) and those that provide different information.

But the goal of PCA is also to identify if there are groups of observations where each group differ from the other groups of observations and to identify the variables that create these groups

3. Covariance PCA

3.1. A first look at the "Beer" data

For our example of a covariance PCA, we use data originally coming from a sensory profile on beers. These data (presented in Table 3.1) give the perceived intensity of twelve attributes used to describe six beers—the intensity is given on a scale going from 0 to 7. Table 3.1 contains the average of the intensity computed for ten panelists who participated to the experiment and Table 3.2 gives the basic statistics.

Table 3.1. The Beer Tasting Data Set

	Bitter	Acid	Sweet	Astringent	Alcoholic	Hop	Cereal	Toasted	Yeast	Hay	Malt	Fermented
Bruxelles	3.2	2.2	2.0	2.5	3.2	3.8	3.3	2.9	3.3	2.8	3.7	3.3
FranzDunkel	3.9	2.1	2.7	3.4	4.3	4.5	3.7	3.0	3.4	3.0	3.7	3.5
FranzNatür	4.3	2.1	3.1	3.4	4.4	4.5	4.1	3.9	3.5	3.3	3.9	2.8
Hofbräu	4.4	3.1	0.0	3.2	4.0	4.7	3.7	3.4	3.6	3.6	4.0	3.7
Paulaner	3.7	2.4	1.0	3.3	3.9	4.7	3.8	3.1	3.4	3.3	3.8	3.3
BlueMoon	3.2	3.3	4.6	3.3	5.2	5.2	2.5	2.5	2.1	2.9	4.2	4.1

Table 3.2. *Basic Statistics for The Beer Tasting Data Set*

	Bitter	Acid	Sweet	Astringent	Alcoholic	Hop	Cereal	Toasted	Yeast	Hay	Malt	Fermented
Mean	3.78	2.54	2.23	3.19	4.18	4.58	3.53	3.13	3.23	3.13	3.89	3.46
SD	0.52	0.52	1.61	0.36	0.65	0.45	0.57	0.48	0.56	0.31	0.20	0.44
Var	0.27	0.27	2.61	0.13	0.42	0.20	0.32	0.23	0.32	0.10	0.04	0.19
SS	1.34	1.37	13.03	0.64	2.11	1.02	1.60	1.16	1.60	0.48	0.21	0.97

Table 3.3. The Centered Data

	Bitter	Acid	Sweet	Astringent	Alcoholic	Hop	Cereal	Toasted	Yeast	Hay	Malt	Fermented
Bruxelles	-0.59	-0.29	-0.23	-0.70	-0.94	-0.76	-0.19	-0.27	0.07	-0.36	-0.20	-0.14
FranzDunkel	0.12	-0.42	0.46	0.23	0.14	-0.10	0.19	-0.17	0.19	-0.16	-0.20	0.05
FranzNatür	0.50	-0.46	0.85	0.24	0.23	-0.04	0.61	0.77	0.26	0.13	0.05	-0.65
Hofbräu	0.62	0.55	-2.23	-0.03	-0.19	0.10	0.17	0.27	0.40	0.48	0.13	0.27
Paulaner	-0.08	-0.13	-1.21	0.14	-0.27	0.17	0.25	0.02	0.21	0.15	-0.10	-0.18
BlueMoon	-0.58	0.76	2.37	0.11	1.02	0.62	-1.03	-0.63	-1.13	-0.23	0.31	0.64

In this example, the intensity of all the variables is measured with the same scale (i.e., from 0 to 7) and therefore we do not want to normalize these variables because we want the differences in intensity to be taken into account by the analysis. By contrast, because we are interested in the difference of intensity between the products, we will center the variables (i.e., subtract the mean of each variable from the scores for this variable). So, the data table that we are, in fact, analyzing is the centered data table given in Table 3.3. This table is not easy to read (because human beings are not very good at dealing with numbers with too many

digits) and so Table 3.3 is easier to understand using its graphical representation (given in Figure 3.1)

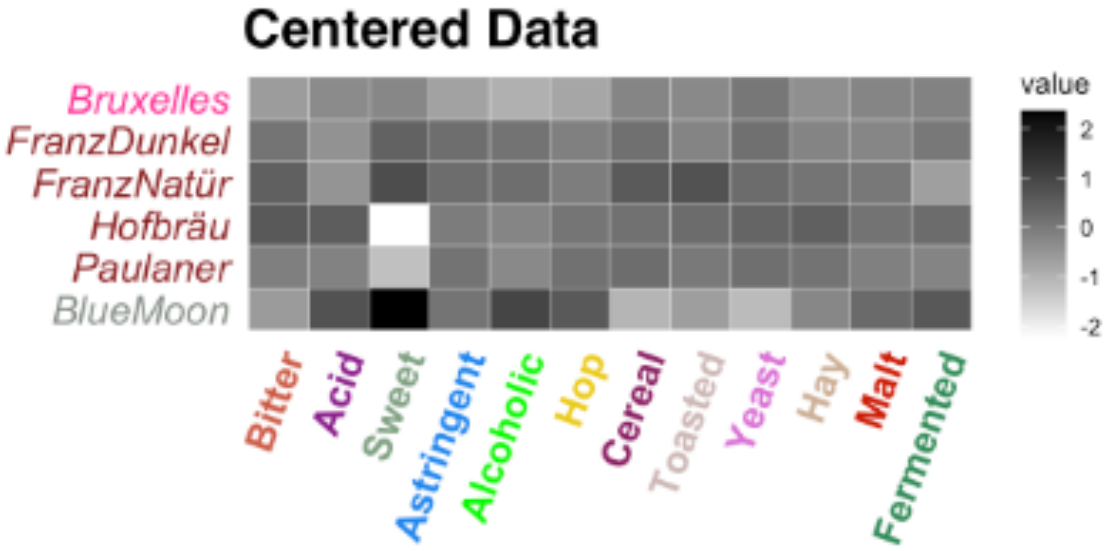

Figure 3.1 The Original Data (Centered).

Before we start principal component analysis, we need to explore the structure of the variables used in this study. Here, because we have decided to center but not to normalize the variables, the relationship between two variables is expressed by their covariance

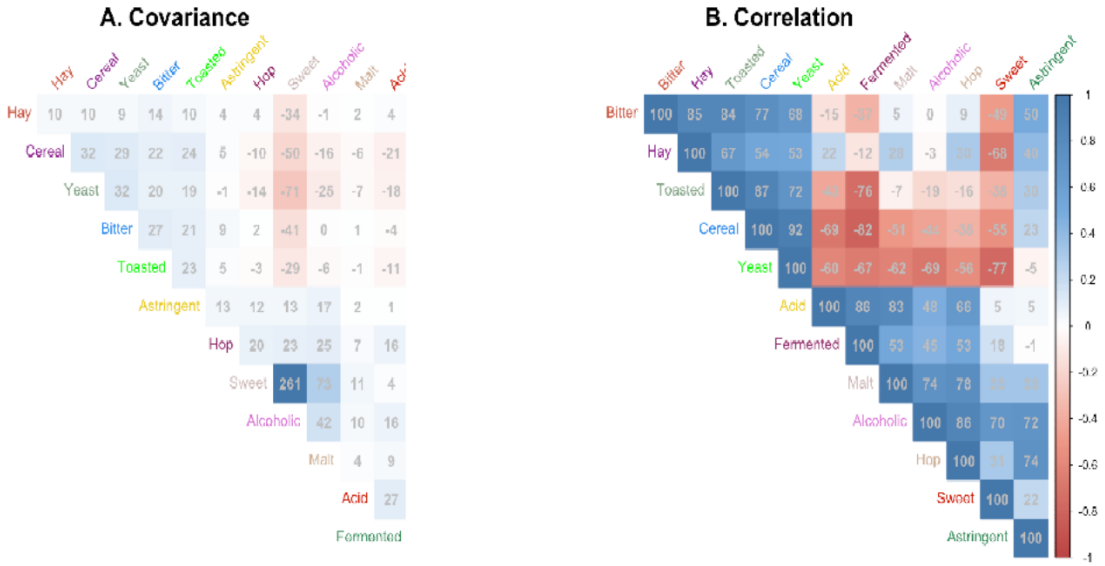

Figure 3.2 .**A** Covariance Matrix HeatMap, **B:** Correlation Matrix HeatMap. Note: The numbers in the heatmaps are multiplied by 100 and rounded.

To better understand the covariance structure of the variables, we can look at a map (called a *heatmap*) such as Figure 3.2A that shows the covariance (multiplied by one hundred and rounded) between variables (note that the diagonal shows the variance of the variables). In the map, for example, we see that *Sweet* (which has by far the largest variance) has large covariances with all the other variables (this is a consequence of having a large variance because the covariance between two variables is equal to their correlation times their standard deviations). So, if we want to understand the pattern of variability in the data, we can expect that the variable Sweet will play an important role. To better understand the pattern of association between variables, we have also added their correlation matrix heatmap (see Figure 3.2B).

3.2. So, what does PCA do?

To identify the most important sources of variability in the data, PCA creates new variables called *components* (also called *factors* or *latent variables*) that are obtained by first assigning a weight to each variable, multiplying each variable by its weight, and then adding these weighted variables together (in technical jargon, we call this process "creating a linear combination" of the variables). The weights are chosen so as to create the largest possible value of the sum of the squared covariances between the *component* and the original variables. The component with the largest sum of the squared covariances is called the *first* component. Often, we will compute several components that will be ordered by this sum of squared covariances (if you feel that "sum of squared covariance" is a long winded name, we will see later on that the official name is in fact "squared eigenvalues," which one do you prefer?). For the second component, we use the same process as for the first component but we add the constraint that this second component is uncorrelated with the first component. We keep this process as long as these components exist (the largest number of such possible components is the smallest of the number of variables or observations minus one). For example, the weights for the first component (given in Table 3.5) are used to compute the first component (denoted as $\mathbf{f}_1$) as:

$$\begin{aligned}\mathbf{f}_1 = {} & .158 \times \text{Bitter} - .074 \times \text{Acid} - .849 \times \text{Sweet} - .040 \times \text{Astringent} \\ & - .266 \times \text{Alcoholic} - .109 \times \text{Hop} + .217 \times \text{Cereal} + .137 \times \text{Toasted} \\ & + .274 \times \text{Yeast} + .109 \times \text{Hay} - .050 \times \text{Malt} - .092 \times \text{Fermented} \ .\end{aligned}$$

Here the variables are centered (see Table 3.3) and all the elements of a variable are multiplied by their weights. So $\mathbf{f_1}$ is computed as:

$$\mathbf{f}_1 = .158 \times \begin{bmatrix} -.59 \\ .12 \\ .50 \\ .62 \\ -.08 \\ -.58 \end{bmatrix} + \cdots - .092 \times \begin{bmatrix} -.14 \\ .05 \\ -.65 \\ .27 \\ -.18 \\ .64 \end{bmatrix} = \begin{bmatrix} -.09 \\ .02 \\ .08 \\ .10 \\ -.01 \\ -.09 \end{bmatrix} + \cdots + \begin{bmatrix} .01 \\ -.00 \\ .06 \\ -.03 \\ .02 \\ -.06 \end{bmatrix} = \begin{bmatrix} \text{Bruxelles} & 0.412 \\ \text{FranzDunkel} & -0.317 \\ \text{FranzNatur} & -0.296 \\ \text{Hofbrau} & 2.202 \\ \text{Paulaner} & 1.221 \\ \text{BlueMoon} & -3.222 \end{bmatrix}$$

The weights used for computing a component are also often called *loadings*—but the term could be misleading because several related (but different) quantities are all called *loadings*. In PCA, the larger the magnitude of the weight of a variable, the more important this variable is for creating the component. But how are these weights obtained? If you are not interested in the technical details, you could accept that this involves some kind of magic and skip the rest of this paragraph, if you enjoy these details, keep on reading. These weights are obtained as the solution of a classical optimization problem known as the *eigenvalue* decomposition or also as the *singular value decomposition*. This optimization problem can be interpreted in several different but equivalent ways. For example, the first component has the property that the sum of the squares of the cross-products of this first component with all the original variables is as large as possible. Here the sum of squares of these cross products between the first component and all the original variables is computed as:

$$7.30 + 1.58 + 210.42 + 0.48 + 20.65 + 3.43 + 13.71 + 5.51 + 21.87 + 3.46 + 0.73 + 2.44 = 291.59.$$

The square root of that sum is equal to $\sqrt{291.59} = 17.075$: it is called the first *eigenvalue* (eigen in the German which means *specific, proper,* or *characteristic*, and, incidentally, old texts often use the term *proper* or *characteristic* value). Eigenvalues are traditionally denoted by the Greek letter λ (lambda, the letter equivalent to the lowercase letter ℓ in our alphabet). So here, we would write that, in PCA, the eigenvalue of a component is also equal to the sum of the squares of all the elements of this component. For example, we could also compute the first eigenvalue as:

$$\lambda_1 = 0.412^2 + (-0.317)^2 + (-0.296)^2 + 2.202^2 + 1.221^2 + (-3.222)^2 = 17.075\,.$$

Finally, the square root of an eigenvalue is called a *singular value* (hence the name *singular value decomposition*) and is traditionally denoted by the lowercase Greek letter δ ("delta"). For example, the first singular value is equal to:

$$\delta_1 = \sqrt{\lambda_1} = \sqrt{17.075} = 4.132 .$$

So, now that we have the first component and the weights to compute it, we need to compute the subsequent components. To find the other components, we subtract from each variable the first component weighted by the weight of this variable. For example, to handle the variable *Bitter* (whose weight was equal to .158) we multiply the first component by .158, and then we subtract it from the variable *Bitter*. After this subtraction, the variable *Bitter* becomes (see Table 3.3 for the values of all the variables and Table 3.5 for the values of the weights for the first component):

$$\begin{bmatrix} \text{Bruxelles} & -0.587 \\ \text{FranzDunkel} & 0.122 \\ \text{FranzNatur} & 0.501 \\ \text{Hofbrau} & 0.622 \\ \text{Paulaner} & -0.080 \\ \text{BlueMoon} & -0.577 \end{bmatrix} - \left(0.158 \times \begin{bmatrix} -0.412 \\ -0.317 \\ -0.296 \\ -2.202 \\ -1.221 \\ -3.222 \end{bmatrix} \right) = \begin{bmatrix} -0.587 \\ 0.122 \\ 0.501 \\ 0.622 \\ -0.080 \\ -0.577 \end{bmatrix} - \begin{bmatrix} 0.065 \\ -0.050 \\ -0.047 \\ 0.349 \\ 0.193 \\ -0.510 \end{bmatrix} = \begin{bmatrix} -0.652 \\ 0.172 \\ 0.547 \\ 0.273 \\ -0.273 \\ -0.067 \end{bmatrix}$$

When applied to all variables from Table 3.3, this procedure gives the following new data table (which is called the deflated table or the residual data table, note that values are rounded to two decimals):

Beer												
1	−0.65	−0.26	0.12	−0.68	−0.83	−0.72	−0.28	−0.32	−0.04	−0.40	−0.18	−0.10
2	0.17	−0.45	0.19	0.22	0.06	−0.13	0.26	−0.12	0.27	−0.13	−0.22	0.02
3	0.55	−0.49	0.60	0.23	0.15	−0.07	0.67	0.81	0.34	0.16	0.04	−0.68
4	0.27	0.71	−0.36	0.06	0.39	0.34	−0.30	−0.03	−0.21	0.24	0.25	0.48
5	−0.27	−0.05	−0.17	0.19	0.06	0.31	−0.01	−0.15	−0.12	0.01	−0.04	−0.06
6	−0.07	0.52	−0.37	−0.02	0.17	0.28	−0.33	−0.18	−0.25	0.12	0.15	0.35

In this deflated data table, the first component having been subtracted from all the variables is eliminated from the data. Then, the procedure used to find the first principal component is applied to the deflated data table and will give the second principal component. This procedure will then give the following weights to create the second component:

$$-0.148; 0.505; -0.403; -0.062; 0.100; 0.217; -0.393; -0.304; -0.257; 0.058; 0.108; 0.415.$$

As a (nice) practical consequence of the way the deflated data table is built (i.e., by subtraction) the second principal component can be computed using the weights indicated above, either with the deflated data table or with the original centered data table (which, in general, is easier for computation). This way, the second principal component, denoted $\mathbf{f}_2$ is computed (from the original variables) as:

$$\begin{aligned}\mathbf{f}_1 = &-.148 \times \text{Bitter} + .505 \times \text{Acid} - .403 \times \text{Sweet} - .062 \times \text{Astringent} \\ &+ .100 \times \text{Alcoholic} + .217 \times \text{Hop} - .393 \times \text{Cereal} - .304 \times \text{Toasted} \\ &- .257 \times \text{Yeast} + .058 \times \text{Hay} - .108 \times \text{Malt} - .415 \times \text{Fermented}\ .\end{aligned}$$

This gives the following values for the second component:

$$\begin{bmatrix} \text{Bruxelles} & -0.146 \\ \text{FranzDunkel} & -0.518 \\ \text{FranzNatur} & -1.448 \\ \text{Hofbrau} & 0.994 \\ \text{Paulaner} & 0.198 \\ \text{BlueMoon} & 0.920 \end{bmatrix}.$$

The process of deflating the data table and computing the principal components one at a time continues till the deflated data table contains only zeros—which, for this data set, happens after five iterations (why five? Because, remember, the number of components cannot be larger than the smallest of the number of columns or rows minus one, and this number here is equal to five). Incidentally, the number of components of a data table is called its *rank*. With this procedure we obtain Table 3.4.

Table 3.4. *The Five Principal Components for the Six Beers*

	f_1	f_2	f_3	f_4	f_5
Bruxelles	0.412	-0.146	1.580	-0.209	0.018
FranzDunkel	-0.317	-0.518	-0.014	0.436	0.306
FranzNatür	-0.296	-1.448	-0.711	-0.280	-0.079
Hofbräu	2.202	0.994	-0.637	-0.217	0.161
Paulaner	1.221	0.198	-0.038	0.334	-0.359
BlueMoon	-3.222	0.920	-0.181	-0.065	-0.046
Sum of Squares (Eigenvalue)	*17.076*	*4.259*	*3.443*	*0.475*	*0.257*

Table 3.4 describes the relationships between the observations (i.e., the beers): The information given in Table 3.4 is, in fact, the same as the information given in the original data table (i.e., Table 3.3) but the information in Table 3.4 is organized such that the largest possible amount of information is provided by the first component, and then of the information left out by the first component, the largest possible amount is given by the second component, and so on till the last component. This pattern guarantees that we have the largest amount of information for a given number of components. For example, Component 1 identifies two beers (i.e., *Hofbräu* and *BlueMoon*) with large magnitudes but opposite signs (2.202 vs. −3.222). Because these two beers have different signs, we can conclude that these are the two most different beers in the data set (a conclusion confirmed by examining the centered data set in Table 3.3 and Figure 3.1). The last row of Table 3.4 gives the eigenvalue of this component obtained here as the sum of squares of its values (and this is another way of computing the eigenvalues of the component). Components with large differences in their values (and so large eigenvalues) differentiate the observations better than components with small differences between their observations (and so small eigenvalues); and, therefore, the eigenvalue of a component measures the information given by a component.

To sum up: components with large eigenvalues carry more information than components with small eigenvalues.

Table 3.5. The Variable Weights for the Principal Components

	f_1	f_2	f_3	f_4	f_5
Bitter	0.158	-0.148	-0.457	-0.186	0.554
Acid	-0.074	0.505	-0.175	-0.438	0.014
Sweet	-0.849	-0.403	0.018	-0.135	0.122
Astringent	-0.040	-0.062	-0.374	0.476	-0.079
Alcoholic	-0.266	0.100	-0.495	0.164	0.095
Hop	-0.109	0.217	-0.394	0.261	-0.448
Cereal	0.217	-0.393	-0.194	0.139	-0.038
Toasted	0.137	-0.304	-0.298	-0.513	-0.199
Yeast	0.274	-0.257	-0.038	0.108	0.304
Hay	0.109	0.058	-0.269	-0.154	-0.117
Malt	-0.050	0.108	-0.140	-0.304	-0.108
Fermented	-0.092	0.415	-0.014	0.153	0.550
Sum of Squares	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*

Each component describes the relationships between the observations, and this implies that to understand these relationships we also need to understand how the variables are combined to create the components. This can be done by looking at the weights applied to the variables. These weights are given in Table 3.5 where we see that *Sweet* has the largest negative weight

(i.e., –0.849) followed by *Alcoholic* (i.e., –0.266). This pattern tells that these two variables create (for the first component and likely for the whole data set) the largest differences among the beers. Note that in Table 3.5, the weights are normalized such that the sum of the squared weights for a component is equal to one. This normalization facilitates the comparison of the weights within one component (e.g., *Sweet*, with a value of –0.849 is the most important variable for creating Component 1 whereas *Acidic* with a value of 0.505 is the most important variable for creating Component 2). But this normalization does not allow to compare importance across dimensions because the weights do not take into account the eigenvalue of the component. To palliate this problem, we can re-normalize the table of weights so that their sum of squares per component is now equal to the component eigenvalue (as shown in Table 3.6). In this table, the weights reflect the importance of a variable for the whole analysis rather than for their component only (e.g., *Sweet*, with a value of −3.51 on Component 1 is the most important variable for the whole analysis whereas *Yeast* with a value of 1.13 on Component 1 is the second most important variable for the whole analysis).

Table 3.6. The Variable Weights (Scaled to Eigenvalues) for the Principal Components

	f_1	f_2	f_3	f_4	f_5
Bitter	0.654	-0.306	-0.848	-0.128	0.281
Acid	-0.304	1.042	-0.325	-0.302	0.007
Sweet	-3.510	-0.832	0.034	-0.093	0.062
Astringent	-0.167	-0.127	-0.694	0.328	-0.040
Alcoholic	-1.100	0.207	-0.918	0.113	0.048
Hop	-0.448	0.448	-0.731	0.180	-0.227
Cereal	0.896	-0.810	-0.360	0.096	-0.019
Toasted	0.568	-0.626	-0.553	-0.353	-0.101
Yeast	1.132	-0.530	-0.071	0.074	0.154
Hay	0.450	0.119	-0.499	-0.106	-0.060
Malt	-0.207	0.223	-0.259	-0.209	-0.055
Fermented	-0.378	0.857	-0.026	0.105	0.279
Sum of Squares (Eigenvalues)	*17.076*	*4.259*	*3.443*	*0.475*	*0.257*

Tables 3.4 and 3.5 (or Table 3.6) together contain all the information needed for PCA but they are not very easy to interpret. To facilitate the interpretation, graphs and numerical indices (called *interpretation helpers*) are derived and are presented below.

3.2.1. How to choose the important components

The first step in interpreting PCA is to identify the important or relevant components. To do so, recall that the components are ordered by their eigenvalues (which express the information provided by the components). Remember, also, that the components are uncorrelated and therefore that their sums of squares can be added together (this is just a consequence of the Pythagorean theorem). The sum of all the eigenvalues is called the total Inertia of the data table (often denoted $\mathcal{I}$, i.e., the uppercase letter *I* in a funny font) and is equal here to 25.51. Moreover, this inertia can also be computed from the original variables as illustrated by Tables 3.2 and 3.3 (and, of course, Tables 3.4 and 3.6).

Table 3.7. The Decomposition of the Total Inertia by the Variables

	Bitter	Acid	Sweet	Astringent	Alcoholic	Hop	Cereal	Toasted	Yeast	Hay	Malt	Fermented	total
Inertia (Sum of Squares)	1.34	1.37	13.03	0.64	2.11	1.02	1.60	1.16	1.60	0.48	0.21	0.97	25.51
Percentage of Inertia	5.24	5.39	51.08	2.49	8.27	4.00	6.27	4.53	6.26	1.88	0.81	3.79	100.00

Table 3.8. The Decomposition of the Total Inertia by the Components

	f_1	f_2	f_3	f_4	f_5
λ: Eigenvalue	17.076	4.259	3.443	0.475	0.257
λ: Cumulative Sum	17.076	21.335	24.778	25.253	25.510
τ: Percentage of Inertia	66.938	16.697	13.497	1.861	1.007
τ: Cumulative Percentage	66.938	83.635	97.132	98.993	100.000

The inertia of the table of data (which is here equal to 25.51) quantifies the variability of these data (cf., Tables 3.7 3.8). The most important variable (by far) is *Sweet* with a sum of squares equal to 13.03—a value that corresponds to 51% (i.e., 13.03/25.01 = .51) of the total inertia. And so, we will say that the variable *Sweet* explains 51% of the total inertia; often we will also say that Sweet explains 51% of the total variance or that it contributes to 51% of the total inertia. Remember that, in PCA, a component is a new variable whose sum of squares (i.e., its eigenvalue) is as large as possible (after the previous components have been found).

Also, because the eigenvalues can be added, the first two components together explain an inertia equals to 17.97 + 4.26 = 21.34—a value that corresponds to 84% of the total inertia (see Table 3.8). Of note, for the same number of variables PCA always explains a larger proportion of inertia than the same number of original variables. For example, with only one component, PCA explains 65% of the total Inertia versus 51% for the variable *Sweet*; and with two components, PCA explains 84% of the total inertia whereas the best two variables (*Sweet* and *Alcoholic*) together explain only 51% + 8% = 69% of the total inertia.

To sum up, the eigenvalue and the percentage of inertia explained by a component represent its importance.

But how to decide if a component must be kept for the analysis? There is no real consensus here, but, as a rule of thumb, we often consider that in order to keep a component, its eigenvalue needs to be larger than the average Inertia. For our example, this rule is equivalent to saying that a component should have an eigenvalue larger than 25.51 / 5 = 5.10 (or, equivalently, that a component must explain at least 20% of the total inertia when we have five variables). This rule is often called the *Kaiser's rule* or *Kaiser's criterion* (from the name of the statistician who first suggested it). This criterion is illustrated in Figure 3.3 that plots the eigenvalues as a function of their rank. In this figure, we also plotted the Kaiser's line (i.e., the line representing the average inertia). According to Figure 3.3, only the eigenvalue of the first component is above the Kaiser's line and, therefore, Kaiser's criterion indicates that we should keep only the first component. Another popular rule is called the "elbow test" or "elbow criterion." Here we look at the difference between two consecutive eigenvalues and identify the place where the difference between these two consecutive eigenvalues is small. At this place the graph looks like an elbow (well, kind of!). In Figure 3.3 the elbow can be seen after the second component and so the elbow rule will suggest to keep the first two components for further inspection. However, note that some authors suggest to keep only the components before the elbow, and, so, these authors would keep only the first component for further inspection. With our example, we will keep the first two components for further inspection.

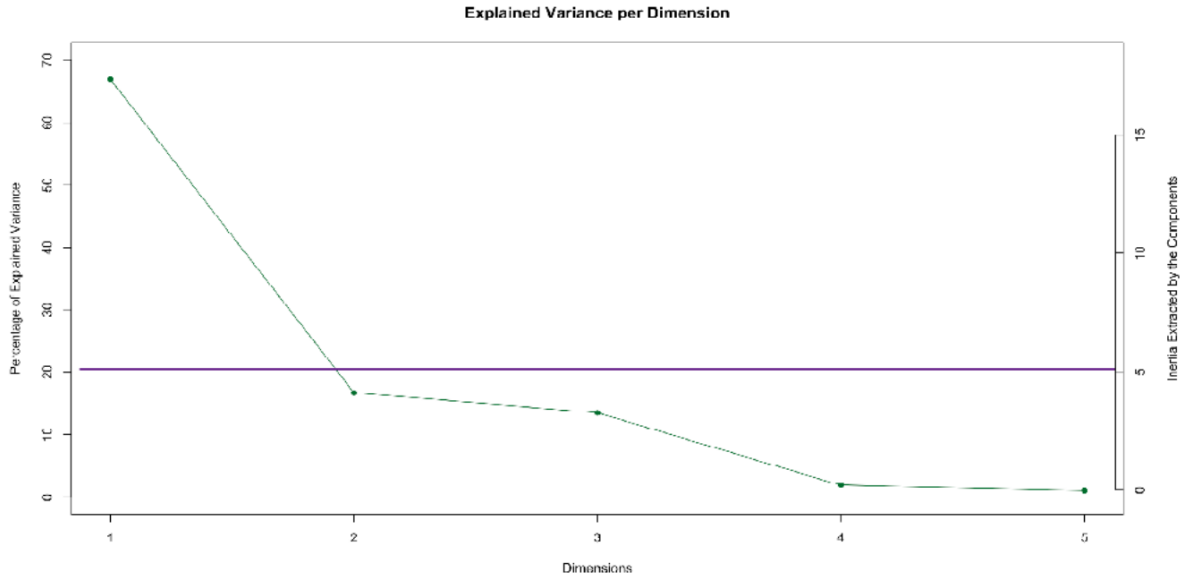

Figure 3.3 The Scree Plot of the Eigenvalues. The Kaiser's line represents the average eigenvalue: Components above the Kaiserline are kept.

Note incidentally, that for sensory profiles realized by trained panelists, a tradition stipulates (but without any real justification) that in order for a PCA to be valid, the percentage of inertia explained by the first two components must be at least 60%.

3.2.2. How to interpret the relationship between variables and components?

3.2.2.1 Understanding the components: weights and contributions

The first step of the analysis is to understand the components (which, remember, are obtained by first multiplying the original variables by their weights and then adding them, see, for example, Table 3.5). In order to create this map, we use the horizontal axis for Component 1, the vertical axis for Component 2, and we use the weights of the variables as their coordinates to plot a point representing the variables by a point. We often add an arrow going from the origin of the map to the point representing the variable. For example, the weights for variable Sweet for Components 1 and 2 are equal to (respectively) –0.849 and –0.403. These numbers are the coordinates for the tip of the arrow representing the variable *Sweet* as shown in Figure 3.4.

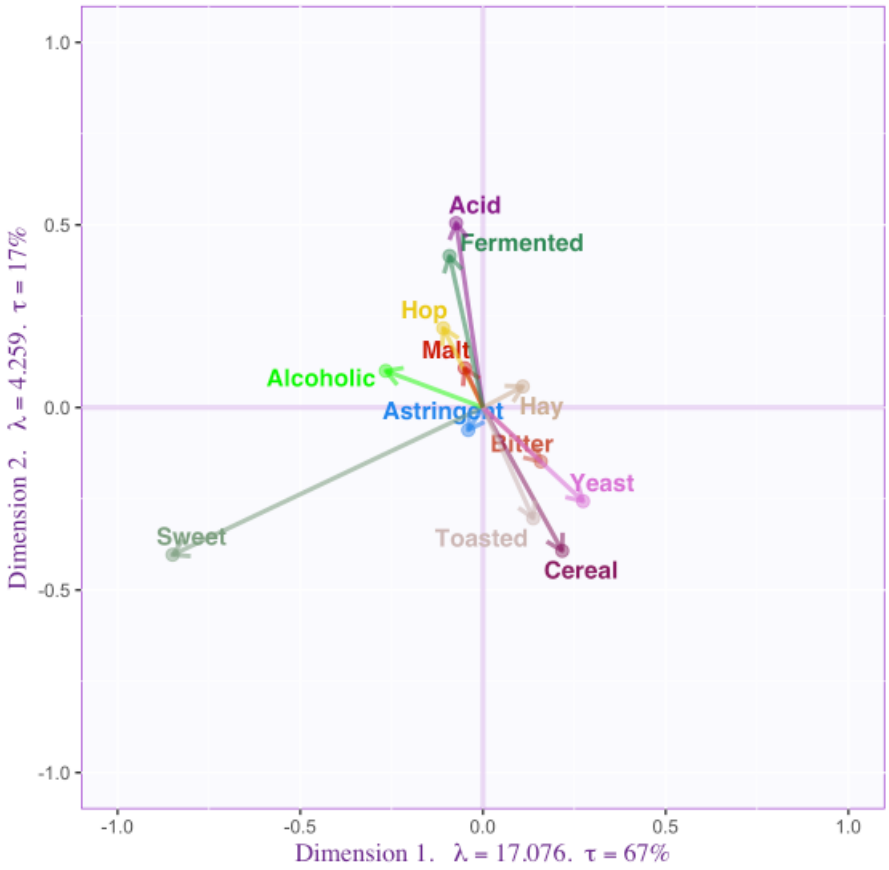

Figure 3.4 Plot of the Variable Weights for Components 1 and 2. The Coordinates of the Tip of the Arrows are the Weights of the Variables.

Figure 3.4 shows that the first component is almost perfectly determined by the variable *Sweet*. This interpretation is confirmed by looking at the values of the weights for the first dimension. To make that examination easier we often re-express these weights as a proportion. To do so, we square each weight and now the sum of the squared weights of all the variables for a given dimension is equal to one (again we use squared values because this is a consequence of the Pythagorean theorem: squared values can be added). This squared weight is called the contribution of a variable to a component.

To sum up: Because a contribution is a proportion, the sum of the contributions of all the variables for a given component is equal to one.

There are two criteria to identify the important components from their contributions: The first criterion considers that a variable is a relevant contributor for a component if its contribution to the component is larger than its contribution to the total Inertia. For example, from Table 3.5, we find that *Sweet* contributes 72% to the Inertia of Component 1, and 51% to the whole data set. With this rule, *Sweet* is an important variable for Component 1 because 72% is larger

than 51%. The second criterion considers that a variable is a relevant contributor for a component if its contribution to the component is larger than the average variable. For example, from Table 3.5, we find that *Sweet* contributes for 72% to the Inertia of Component 1. With this rule, *Sweet* is an important variable for Component 1 because 72% is larger than the average contribution which is equal to 100 / 12 = 8%. As this example shows, these two criteria could differ drastically, but in practice, they often concur.

The contributions of the variables to the components are given in Table 3.9 where the positive contributions are in green, the negative contributions in red, and the important contributions in bold italic. For example, this table confirms (again) that *Sweet* is the major contributor to Component 1 because it contributes to 72% of this component.

A graphical way to easily identify the important contributions is to plot the signed contributions for a given dimension as a histogram. For example, Figure 3.5 displays the contributions for Components 1 and 2. In this figure, relevant contributions are colored and irrelevant contributions are grayed (using the second criterion for relevance). These graphs confirm what the table of contributions suggested: the variable *Sweet* is the only important contributor for Component 1. By contrast, Component 2 is mostly created, on one hand by *Fermented* and *Acid* (with positive weights) and, on the other hand, by *Cereal* and *Sweet* (with negative weights).

Table 3.9. The Contributions of the Variables to the Components (Expressed as Percentages). Positive Contributions are in Green, Negative Contributions are in Red. Contributions Larger than Average (i.e.,100 / 12 = 8.3) are Typeset in Bold Italic

	f1	f2	f3	f4	f5
Bitter	3	-2	***-21***	-3	***31***
Acid	-1	***25***	-3	***-19***	0
Sweet	***-72***	***-16***	0	-2	1
Astringent	0	0	***-14***	***23***	-1
Alcoholic	-7	1	***-24***	3	1
Hop	-1	5	***-16***	7	***-20***
Cereal	5	***-15***	-4	2	0
Toasted	2	-9	-9	***-26***	-4
Yeast	8	-7	0	1	9
Hay	1	0	-7	-2	-1
Malt	0	1	-2	-9	-1
Fermented	-1	***17***	0	2	***30***

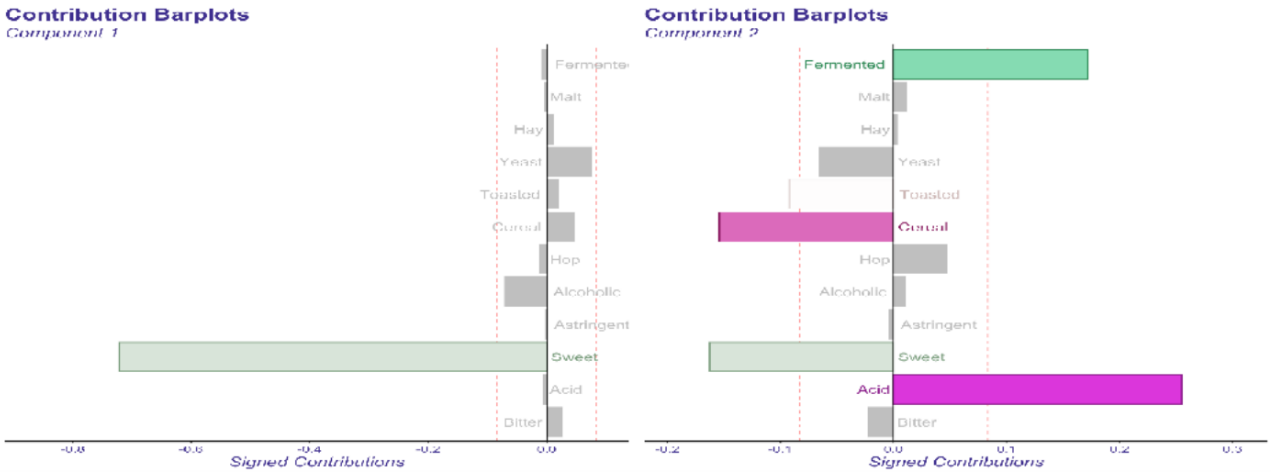

Figure 3.5 Histogram of the Contributions of The Variables

3.2.2.2 Understanding the Variables from the Components

We have just seen that to understand a component we need to look at the weights of the variables. This way, for example, we could conclude that the first component was mostly

created by the variable *Sweet.* But, by contrast, to understand how the components explain the variables, we need to use the inverse approach and compute the correlations between the components and the variables (as if we wanted to predict the variables from the components). These correlations are shown in Table 3.10. Here, for example, we see that the variables *Sweet* and *Yeast* are strongly correlated with the first component (with, respectively, values of –.97 and .90) and, so, we can say that these two variables are well explained by the first component. Along the same lines, the variables *Acidic* and *Fermented* are strongly correlated with the second component (with, respectively, values equal to .89 and .87): These variables are well explained by the second component.

Note that the sum of the squared correlations of a variable across the components is always equal to one. Once again, this is a consequence of the Pythagorean theorem (squared quantities are additive). Because the sum of the squared correlations for one variable is equal to one, an alternative way of looking at the correlations is to use the squared correlations as shown in Table 3.11. Since a squared correlation can be interpreted as a proportion of common variance, we can interpret the squared correlations as proportions of explained variance (of a variable explained by a component). For example, from Table 3.11 we find that Component 1 explains 94.58% of the variance of *Sweet* and we conclude that almost all the variance of *Sweet* is explained by the first component.

A good way to have a global vision of the structure of the relationships between variables and components is to create a map called the circle of correlation. In this map, the coordinates of a variable are the values of the correlations between this variable and the components that are represented by the horizontal and vertical dimensions of the map. For example, the position of *Bitter* on the map is indicated by the values of .57 and –.26 (see Figure 3.6). Traditionally the position of the variable on this map is indicated by an arrow going from the origin of the map to the position of the variable (in a way similar to the representation of the variables in the weight map).

Because the components are linear combinations of the original variables, for every variable in the data set, the sum of its squared correlations (in the data set) with all the components is always equal to one. This way, for the particular case of the first two components, the maximum value of the sum of the squared coordinates is equal to one. And, therefore, from the Pythagorean theorem (and standard Cartesian geometry) we know that variables perfectly

explained by the first two components will lay on the perimeter of a circle of radius one (remember a circle is the locus of the points whose sum of the squared coordinates is equal to the radius of the circle). This circle of radius one is called the *"circle of correlations"* and we plot this circle on the correlation map because it gives a scale: The closer a variable is to the circle, the better its variance is explained by the components used to draw the circle. Therefore, the correlation between two variables is approximated by their angle, and the correlation between a variable and a component is also approximated by their angle: A small angle indicates a strong positive correlation (exactly +1 when the angle is equal to zero), an angle close to 90 degree indicates a correlation close to zero (exactly zero, when the angle is equal to 90 degrees), and an angle larger than 90 degrees indicates a negative correlation (exactly –1 when the angle is equal to 180 degrees). The quality of the approximation provided by these angles is larger the closer to the circle the variables (it is exact when the variables are on the circle).

For example the variables *Acidic* and *Fermented* are very close to the circle of correlation and the angle that these two variables form with the origin of the circle is small: therefore we conclude that they are highly correlated (their correlation is equal to .83, a value corresponding to an angle of 34°); but *Acidic* and *Sweet* (that are both close to the circle) form almost a right angle with the origin: These two variables are almost orthogonal (their correlation is equal to .05, a value corresponding to an 87° angle); finally *Acidic* and *Cereal* (both close to the circle) form a large obtuse angle with the origin: these two variables are negatively correlated (their correlation is equal to –.69, a value that corresponds to an angle of 134°). By contrast, the variables *Astringent* and *Bitter* are far from the circle of correlation and the angle that they form with the origin is almost a right angle—a configuration that could suggest that these two variables are uncorrelated—but their correlation is, in fact, equal to .50 (a value corresponding to a 60° angle). This last example illustrates that angles are not good approximations of correlation for variables far from the circle of correlations.

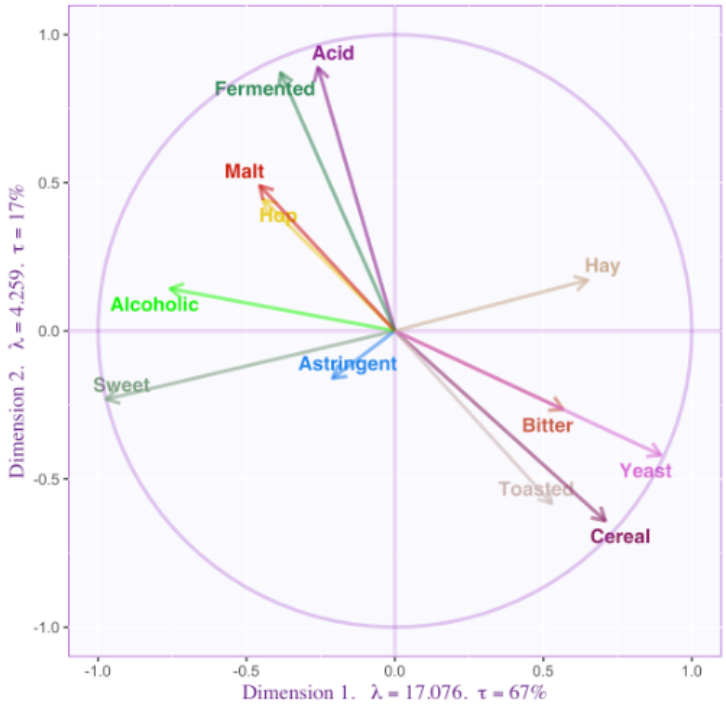

Figure 3.6 Plot of the Circle of Correlations for Components 1 and 2. The Coordinates of the Tip of the Arrows are the Correlations between the Variables and the Components.

The configuration of the correlations (see Tables 3.10 and 3.12, and Figure 3.6) reveals that three components are enough to understand most of the structure of the original variables. For example, the variables *Sweet* and *Yeast* are essentially explained by the first component (with correlations respectively equal to –.97 and .90). This pattern of correlation indicates that sweet beers do not taste like yeast and beers with a taste of yeast are not sweet. Note, again, the difference between the information provided by the weights of the variables and the information provided by the correlations between the variables and the components: the first component is essentially created by the variable *Sweet*, but the first component explains most of the variance of the variables *Sweet* and *Yeast*. This difference between these two conclusions can be attributed to the fact that the data are only centered (i.e., not normalized): When we do not normalize the variables (as is the case here) variables with a large variance (e.g., *Sweet*) contribute more to the creation of the components than variables with a small variance (such as *Yeast*).

Table 3.10. Correlations Between Variables and Principal Components.

	f_1	f_2	f_3	f_4	f_5	SS
Bitter	0.566	-0.265	-0.734	-0.111	0.243	1
Acid	-0.259	0.889	-0.277	-0.257	0.006	1
Sweet	-0.973	-0.231	0.009	-0.026	0.017	1
Astringent	-0.210	-0.160	-0.871	0.412	-0.051	1
Alcoholic	-0.757	0.143	-0.632	0.078	0.033	1
Hop	-0.444	0.444	-0.724	0.178	-0.225	1
Cereal	0.709	-0.641	-0.285	0.076	-0.015	1
Toasted	0.528	-0.583	-0.514	-0.329	-0.094	1
Yeast	0.896	-0.419	-0.056	0.059	0.122	1
Hay	0.649	0.171	-0.720	-0.153	-0.086	1
Malt	-0.455	0.490	-0.571	-0.461	-0.121	1
Fermented	-0.385	0.871	-0.026	0.107	0.284	1

Table 3.11. Squared Correlations Between Variables and Principal Components.

	f_1	f_2	f_3	f_4	f_5	Sum
Bitter	0.320	0.070	0.539	0.012	0.059	1
Acid	0.067	0.790	0.077	0.066	0.000	1
Sweet	0.946	0.053	0.000	0.001	0.000	1
Astringent	0.044	0.026	0.758	0.169	0.003	1
Alcoholic	0.573	0.020	0.399	0.006	0.001	1
Hop	0.197	0.197	0.524	0.032	0.051	1
Cereal	0.502	0.411	0.081	0.006	0.000	1
Toasted	0.279	0.340	0.265	0.108	0.009	1
Yeast	0.803	0.176	0.003	0.003	0.015	1
Hay	0.421	0.029	0.519	0.023	0.007	1
Malt	0.207	0.240	0.326	0.212	0.015	1
Fermented	0.148	0.759	0.001	0.011	0.080	1

Table 3.12. Squared Correlations between Variables and Components (Expressed as Percentages). Square Correlations Larger than Average (e.g., 100 / 5 = 20) are in Bold-Italic.

	f1	f2	f3	f4	f5
Bitter	***32***	7	***54***	1	6
Acid	*7*	***79***	8	7	0
Sweet	***95***	5	0	0	0
Astringent	*4*	3	***76***	17	0
Alcoholic	***57***	2	***40***	1	0
Hop	*20*	20	***52***	3	5
Cereal	***50***	***41***	8	1	0
Toasted	***28***	***34***	***26***	11	1
Yeast	***80***	18	0	0	1
Hay	***42***	3	***52***	2	1
Malt	***21***	***24***	***33***	***21***	1
Fermented	*15*	***76***	0	1	8

But this difference also reflects the difference between the points of view of these two approaches: The weights are used to understand the components from the original variables, whereas the correlations are used to understand the variables from the components. For example, the variables *Acidic* and *Fermented* are mostly explained by the second component (with correlations respectively equal to .89 and .87 and squared correlations equal to .79 and.76). Because the sign of the correlation of these two variables is the same, these two variables concur and so these two variables, being positively correlated, provide the same information. Finally, the variable Astringent is mostly explained by the third component (with a correlation equals to –.87); this third component is also important (but less so) for the variables *Bitter, Hop* and *Hay* see (see Tables 3.10 and 3.12).

3.2.3. Interpreting the Observations from the Components

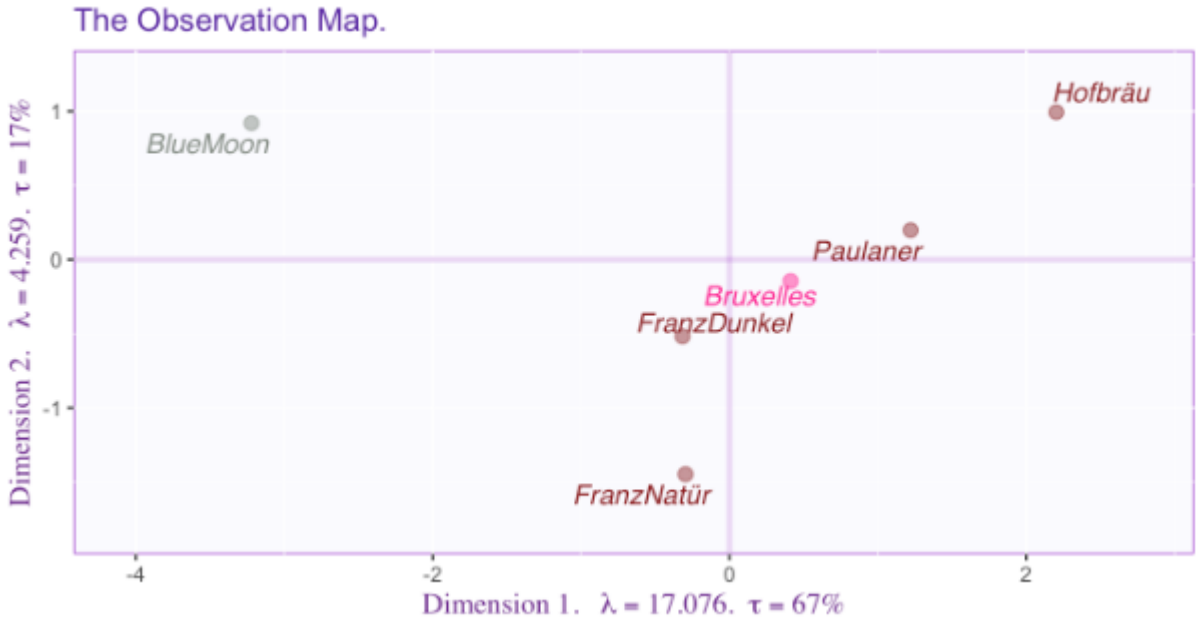

Figure 3.7 Plot of the Observations for Components 1 and 2. The Coordinates of the Points are the Values of the Observations for the Components.

Just like what we did for the analysis of the variables, we use graphical representations and "interpretation helpers" to understand the structure of the observations. The first map uses for coordinates of the observations their values of the component (i.e., as given in Table 3.4, see Figure 3.7) and are called components observations maps (or factor scores maps). In these maps, the relationships between observations are represented by their proximity: observations close to each other on the graph are described, in roughly the same fashion by the variables, whereas observations far from each other are different. Note that the rules used to interpret the observations differ from the rules used to interpret the variables. Observations are

interpreted using the *distance* on the map, but variables are interpreted using the *angle* they make with the origin.

If we happened to have additional information about the observations, we can use it to color the observations on the maps. For example, we know the country of production of the beers, and we will use this information to color the name of the beers on the map: This way, we see right away that the American beer *Blue Moon* is isolated on the first dimension that we previously interpreted as the dimension of the *Sweet* taste). Because the beer *Blue Moon* lays in the same side of the component as variable *Sweet* does (i.e., they both negative), we conclude that the largest source of variance in the observations comes from the opposition of the *Sweet* beer *Blue Moon* to all the other beers (mainly to *Hofbräu*, the least sweet of the beers)—an interpretation confirmed by looking at the original centered data.

3.2.4. Interpretation helpers for the observations

Table 3.13. The decomposition of the Inertia from the observations.

	f_1	f_2	f_3	f_4	f_5	Sum = d2
Bruxelles	0.170	0.021	2.498	0.044	0.000	2.732
FranzDunkel	0.100	0.268	0.000	0.190	0.094	0.652
FranzNatür	0.088	2.097	0.505	0.078	0.006	2.774
Hofbräu	4.850	0.988	0.406	0.047	0.026	6.316
Paulaner	1.490	0.039	0.001	0.112	0.129	1.771
BlueMoon	10.379	0.846	0.033	0.004	0.002	11.264
Sum = eigenvalue	*17.076*	*4.259*	*3.443*	*0.475*	*0.257*	*25.510*

In principal component analysis, the total variability of the data (i.e., the inertia) is decomposed in a parallel fashion by the variables and by the observations, therefore the same indices are used for the observations and for variables when we want to identify the important observations or the important components. Therefore, just like we did for the variables, for the observations we compute contributions (see Table 3.14), correlations (in general called cosines) and squared correlations (called squared cosines, of course, see Table 3.15).

3.3. Contributions for the observations

For the observations, the contributions for a given component are computed with the following procedure (see Table 3.13): 1) square all the values of the observations for this component, and then 2) compute this sum for this component (remember that this sum is equal to the eigenvalue of the component see, Table 3.13), and finally 3) divide each of these squared values by the eigenvalue of the component (see Table 3.14). For example, the contribution (denoted as 'ctr') of the beer *Blue Moon* is computed as:

$$\text{ctr}_{\text{BlueMoon}} = \frac{-3.22^2}{17.08} = \frac{10.38}{17.08} \approx 0.608 = 60.8\% \approx 61\% \,.$$

And just like we did for the variables we can display the contributions with a bar plot as illustrated in Figure 3.8.

Table 3.14. The Contributions of the Observations to the Components (Expressed as Percentages). Positive Contributions are in Green, Negative Contributions are in Red. Contributions Larger than Average (i.e., 17) are Typeset in Bold Italic.

	f1	f2	f3	f4	f5
Bruxelles	1	0	***73***	***-9***	0
FranzDunkel	-1	-6	0	***40***	***36***
FranzNatür	-1	***-49***	***-15***	***-16***	-2
Hofbräu	***28***	***23***	***-12***	***-10***	***10***
Paulaner	***9***	1	0	***24***	***-50***
BlueMoon	***-61***	***20***	-1	-1	-1

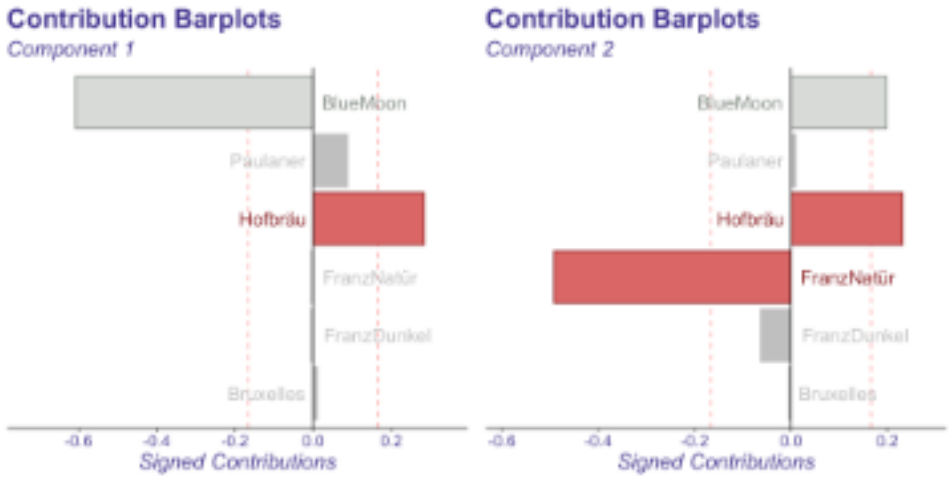

Figure 3.8 Histogram of the Contributions of the Observations

3.3.2. Squared cosines

For the observations, by contrast with the variables, we prefer (just as quirk of history) the appellation cosine (and squared cosines) to the appellation correlation used for the variables. Squared cosines are computed from Table 3.13. The squared cosine of an observation is obtained by dividing its squared coordinate by the row total (which gives the squared Euclidean distance of the observation to the center of the space, which would be the average beer).

The squared cosine of an observation and a component is the squared cosine of the angle made by this observation (in the whole 5-dimensional space) and its projection on this component (see Table 3.13). For example, the distance to the average beer for the beer *Blue Moon* is equal to 11.26, and its coordinate on the first component is equal to 10.38; this gives a squared cosine between *Blue Moon* and the first component equal to 10.38 / 11.26 = .92 = 92%. This high value of the cosine shows that *Blue Moon* is mostly explained by the first component. Therefore, we conclude that the beer *Blue Moon*, being characterized by the first dimension, (which is mostly the taste *Sweet* as indicated in Table 3.9) differs from all the other beers by its strong *Sweet* taste.

Table 3.15. The Squared Cosines (Observations) Expressed as Percentages.

	f_1	f_2	f_3	f_4	f_5	Sum
Bruxelles	6	1	91	2	0	100
FranzDunkel	15	41	0	29	14	100
FranzNatür	3	76	18	3	0	100
Hofbräu	77	16	6	1	0	100
Paulaner	84	2	0	6	7	100
BlueMoon	92	8	0	0	0	100

3.3.2.2 Explaining the Observations: The Circle of Cosines

Just like the correlations for the variables, the sum of the squared cosines of an observation (across the components) is equal to one, and, therefore, we can use the same map for the cosines, by first using the cosines as coordinates and add a circle of Radius one to the observations cosine map (see 3.9). To obtain these cosines we just take the square root of the squared cosines and add the sign of the coordinates (see Table 3.16). The interpretation of the circle of cosines follows the same rules as the circle of correlations: An observation close to the circle is well represented by the components used to draw the circle. By contrast, an observation close to the center of the circle is poorly represented by the components. For example, as seen in Figure 3.9, Blue Moon which lays on the circle is perfectly explained by the first Two components. But, by contrast, *Bruxelles*—being close to the center of the circle—is poorly explained by the first two components—an interpretation confirmed by looking at Tables 3.15 and 3.16 which, together, reveal that *Bruxelles* is mostly explained by the third component of the analysis

Table 3.16. The Cosines between Observations and Components

	f_1	f_2	f_3	f_4	f_5
Bruxelles	0.25	-0.09	0.96	-0.13	0.01
FranzDunkel	-0.39	-0.64	-0.02	0.54	0.38
FranzNatür	-0.18	-0.87	-0.43	-0.17	-0.05
Hofbräu	0.88	0.40	-0.25	-0.09	0.06
Paulaner	0.92	0.15	-0.03	0.25	-0.27
BlueMoon	-0.96	0.27	-0.05	-0.02	-0.01

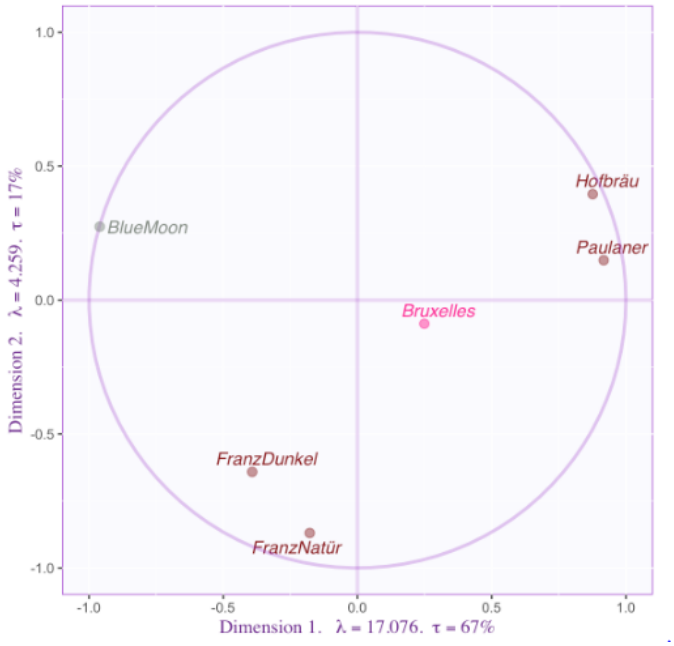

Figure 3.9 Plot of the circle of cosines for the observations and components 1 and 2. The coordinates of the points are the value of the cosine between the observations and the components.

3.3.3. All together now

To sum up: Principal component analysis gives the best view of the data in a space of small dimensions (often displayed as maps) of the pattern of multivariate data sets. To do so, PCA creates new variables called *components* that are obtained by assigning weights to the original variables, multiplying the variables by these weights, and then taking the sum of these weighted variables (a process called creating a linear combination of the variables). To

understand a component, we first need to identify (numerically or graphically) the important variables for this component: those are the variables with a large weight and therefore an important contribution. Once the components and their relationships with the original variables are understood (i.e., using contributions and correlations), we then use the components to analyze and visualized the structure of the observations (again via maps). Note that the rules to interpret the graphs from PCA differ for the variables and the observations. The variables are interpreted using the angles that they form with themselves and the components: A small angle means a large correlation, a right angle means no correlation, and a large angle means a negative correlation. The observations, by contrast, are interpreted using the distance between observations: Close observations are alike, whereas far away observations differ. Finally, to understand the results of a PCA we need to integrate the information given by the variables with the information given by the observations.

However, the interpretation of the relationships between the variables and the observations needs to consider the quality of their representation (i.e., correlations and cosines) on the component or set of components: The larger the magnitude of a cosine or correlation between an observation (respectively a variable) and a component, the better represented is this observation (respectively variable) by this component. In general (but there are exceptions) the observations well represented are those that are the furthest away from the center of the graph (which represents the average observation). In our example the variable *Sweet* is almost completely explained by the first component and the variable *Yeast* is also largely explained by the first dimension. Because their signs on the first component are opposite (*Sweet* is negative and *Yeast* is positive) we conclude that these two variables are negatively correlated (e.g., large values of one variable go with low values for the other variable)—an interpretation confirmed by the important negative correlation equal to –.77. Along the same lines, the beers *Blue Moon, Paulaner*, and *Hofbräu* are well represented on the first dimension and, therefore, their relative position on the circle of cosines (see Figure 3.9) correctly reflects their similarities and differences: *Paulaner*, and *Hofbräu* are alike, and they both differ from *Blue Moon*. By contrast, *Bruxelles* is close to *Paulaner* but is badly represented on the first two components (in fact, the tables of cosines and contributions shows that *Bruxelles* is represented almost completely on the third component), therefore the proximity between *Bruxelles* and *Paulaner* does not imply that they are similar.

3.3.4. All together now again for more: Biplot

Table 3.16. Biplot**.** Observations normalized to the component's singular value.

Observations	$\mathbf{f}_1$	$\mathbf{f}_2$	$\mathbf{f}_3$	$\mathbf{f}_4$	$\mathbf{f}_5$
Bruxelles	0,08	−0,04	0,47	−0,10	0,01
Frans Dunkel	−0,06	−0,15	0,00	0,21	0,18
Frans Natur	−0,06	−0,41	−0,21	−0,14	−0,05
Hofbrau	0,44	0,28	−0,19	−0,11	0,09
Paulaner	0,25	0,06	−0,01	0,16	−0,21
Blue Moon	−0,65	0,26	−0,05	−0,03	−0,03
Somme des carrées	0,69	0,34	0,31	0,11	0,08
Somme des carrés × 12 = δ	4,13	2,06	1,86	0,69	0,51

Table 3.17. Variables normalized to the component's singular value.

Variables	$\mathbf{f}_1$	$\mathbf{f}_2$	$\mathbf{f}_3$	$\mathbf{f}_4$	$\mathbf{f}_5$
Amer	0,09	−0,06	−0,18	−0,04	0,11
Acide	−0,04	0,21	−0,07	−0,10	0,00
Sucré	−0,50	−0,17	0,01	−0,03	0,02
Astringent	−0,02	−0,03	−0,15	0,11	−0,02
Alcool	−0,16	0,04	−0,19	0,04	0,02
Houblon	−0,06	0,09	−0,15	0,06	−0,09
Céréales	0,13	−0,16	−0,08	0,03	−0,01
Grillé	0,08	−0,13	−0,12	−0,12	−0,04
Levure	0,16	−0,11	–0,1	0,03	0,06
Foin	0,06	0,02	−0,11	−0,04	−0,02
Malt	−0,03	0,04	−0,05	−0,07	−0,02
Fermenté	−0,05	0,17	−0,01	0,04	0,11
Somme des carrés	0,34	0,17	0,15	0,06	0,04
Somme des carrés × 6 = δ	4,13	2,06	1,86	0,69	0,51

The graphs of the observations and the variables are traditionally separated, because the interpretation rules differ for these two sets (i.e., angles for the variables and distances for the observations). However, it is possible—especially when the numbers of variables and observations are small—and sometimes convenient to draw variables and observations on the same graph which is then called a *biplot graph* (a term created by Gabriel, 1971; see also Greenacre 2010 for a nice review).

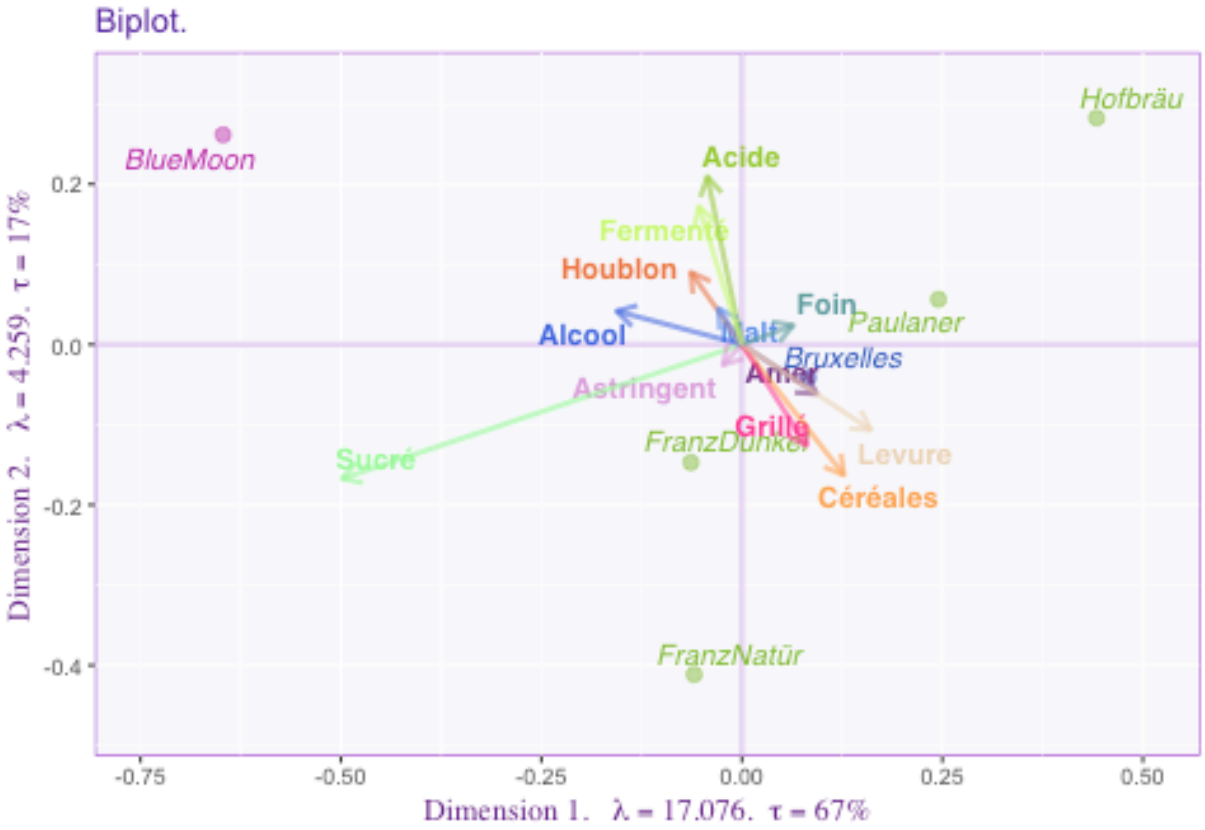

Figure 3.10. *Biplot* for the first two factors of the PCA for he beers example.

To obtain such a simultaneous representation, the first problem to solve is a scaling problem because the variances of the representations differ: The sum of the squares of the weights (and therefore of the variables) for a component is equal to one but the sum of the squares of a component (and therefore of the observations) is equal to its eigenvalue. A first approach to

solve this problem is to re-normalize the weights and components so that these two sets now have the same sum of squares (and this is sometimes the solution chosen). In this case, each set will be normalized so that the sum of the squares of the variables and observations for a component is equal to the singular value of this component. But this does not completely solve the problem because, if the number of variables and observations differ, then the set with the largest number of elements will on average have smaller values than the other set. A better normalization will, therefore, also take into account the number of elements of each set and normalize these sums of squares so that the sum of the squares of a component multiplied by the number of elements considered (i.e., number of observations or variables) is equal to the singular value. The data thus normalized are found (respectively) in Tables 3.16 (for the observations) and 3.17 (for the variables). This way, observations and variables, being commensurable, can now be drawn on the same map as illustrated in Figure 3.10. But how to interpret this simultaneous representation of Figure 3.10? To interpret the relationships between elements of the same set, the usual rule remains valid: observations are points and are interpreted using their distances but variables are lines and are interpreted from their angles, and, in addition, their length represents their variance. The question is therefore: how to interpret the relationships between observations and variables? To do so, we first start by interpreting the center of the map; for the observations this point corresponds to the average beer which therefore has values of 0 for all its variables (because we have centered the data); for the variables this point represents a variable whose values for all the observations are 0. Also the arrow that represent a variable points towards the positive part of the variable. For example, the arrow for the variable *Sweet* (cf. Figure 13A) starts from the value 0 (the average) and points towards the positive values, but this representation forgets the negative values. We can add them as shown in Figure 3.11B which shows in red the line representing the negative values. The relationship between observations (points) and variables (lines) is expressed by the (orthogonal) projection of the observation-point on the variable-line as illustrated in Figure 3.11C which shows the projection of the sweetest beer (*BlueMoon*) on the variable *Sweet*, this observation is projected on the positive end of the arrow representing the variable *Sweet*; similarly, (cf. Figure 3.11D) the least sweet beer (*Hofbrau*) is projected on the negative end of the variable. Finally, we can project all the observations on the variable *Sweet* as shown in Figures 3.11 E and F. The projections of the observations on the variable *Sweet (*Figures 3.11F), give the best approximation in two dimensions of the original values

of the variable *Sweet* as shown by the comparison between the values predicted by the projection of the observations and the measured values given by Table 3.19. The high quality of this prediction is attested by a correlation coefficient with a value of .96 between original and predicted values. Note that the values of the extreme observations (i.e., *BlueMoon* and *Hofbrau*) are better predicted than those of the median observations, because the extreme observations for *Sweet* are almost perfectly represented by the first two components.

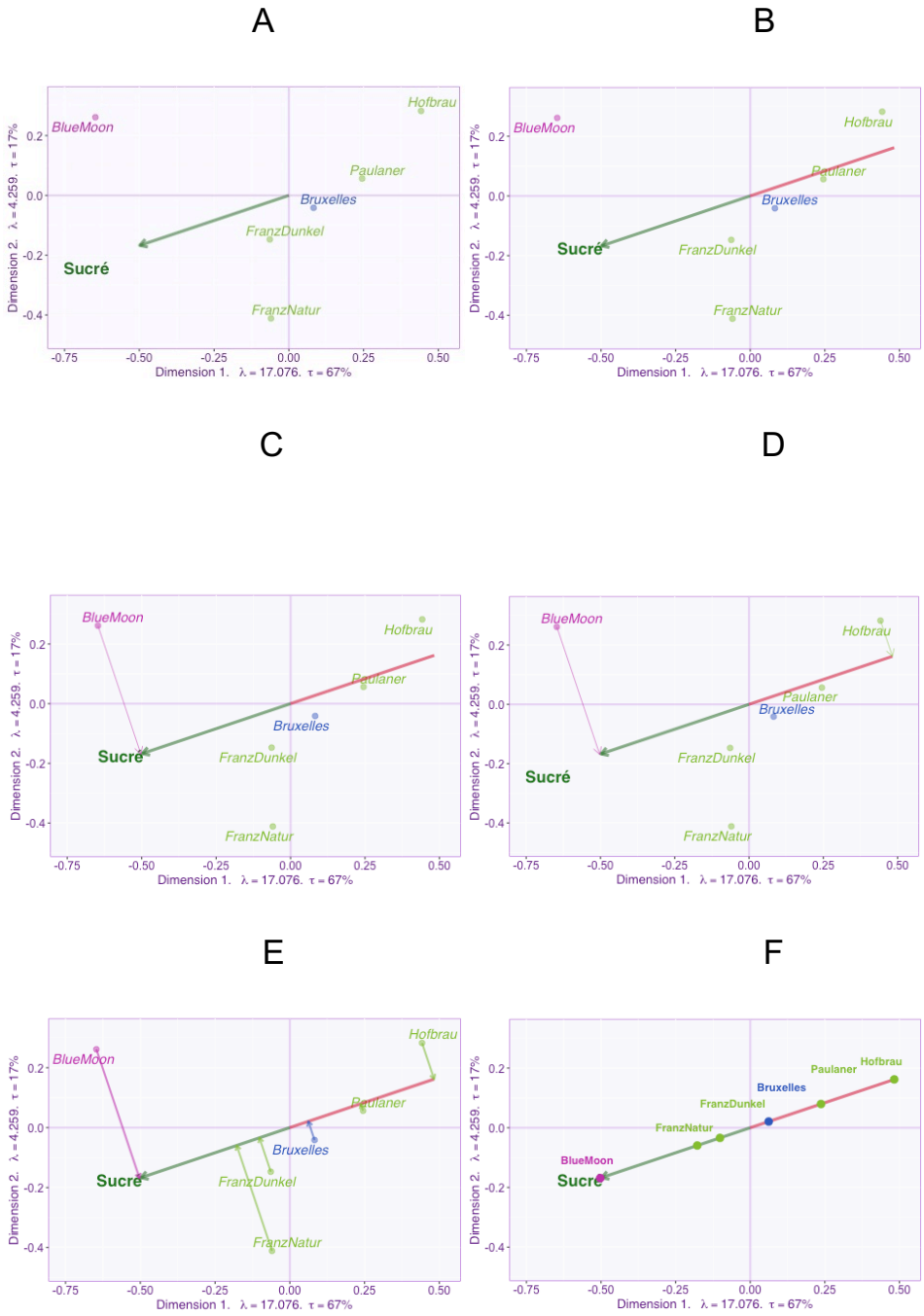

Figure 3.11. *Biplot*. How to interpret a simultaneous representation of observations and variables A) the variable *Sweet* (positive part) and its beers, B) the variable *Sweet* positive and negative parts, C) projection of the beer *Blue Moon* onto the variable *Sweet*, D) projection of *Hofbrau* onto the variable

Sweet, E) projection of all beers on the *Sweet* variable, and F) the projections of the beers onto the *Sweet* variable approximate their values for the *Sweet* variable.

Table 3.17 *Biplot*. Predicted (and actual) values of the variable Sweet. The predicted values are obtained from the projection of the beers onto the variable *Sweet* in the plane created by the first two components of the Biplot. (see Figure 3.11F).

	Biplot	**Sweet Predicted**	**Sweet**
Hofbrau	–0,416	–2,319	–2,234
Paulaner	–0,100	–0,561	–1,206
Bruxelles	–0,007	–0,038	–0,232
FranzDunkel	0,018	0,103	0,456
FranzNatur	0,056	0,314	0,851
BlueMoon	0,451	2,515	2,366

In summary, for biplots, the relationship between observations and variables is represented by the projection of the points representing the observations onto the lines representing the variables. Finally, note that the simultaneous representation of observations and variables is only practical for small data sets.

3.3.5. Supplementary Observations and Variables

Once the analysis is done, new variables and observations can be added to the analysis even though they were not used perform the analysis: These new items are then said to be projected as *supplementary elements* by opposition the variables and observations used for the analysis, which are then called *active* elements. Supplementary elements are also called *supplemental*, *illustrative*, *passive*, or even "*out of sample elements*." The additional observations are described by the same variables as the active observations and the supplementary variables describe the active observations. To project a supplementary observation, the first step is to preprocess it like the active observations, then use the same weights as the active observations to calculate the components. Let's take as an example a beer prototype, which we (i.e., the authors) have evaluated and whose values are given in Table 3.18. To "project as an

supplementary observation" this beer, it is first centered using the average values obtained for the original data. For example, we rated the beer prototype 4.20 for the variable *Bitter* whose average (given in Table 1) is equal to 3.78 for the 6 original beers which gives as centered value for *Bitter*: 4.20 – 3.78 = 0.42 (cf. Table 3.18). To compute the component values, we use the weights that were used for the active observations. This way, the value of the prototype for the first component is computed as

$$\mathbf{f}_{1sup} = 0.158 \times \textit{Bitter} - 0.074 \times \textit{Acid} - 0{,}849 \times \textit{Sweet} - 0.040 \times \textit{Astringent} - 0.266 \times \textit{Alcoholic} - 0.109 \times \textit{Hop} + 0.217 \times \textit{Cereal} + 0.137 \times \textit{Toasted} + 0.274 \times \textit{Yeast} + 0.109 \times \textit{Hay} - 0.50 \times \textit{Malt} - 0.092 \times \textit{Fermented}$$

$$\mathbf{f}_{1sup} = (0.158 \times 0.42) - (-0.074 \times 0.14) - (0.849 \times 1.5) - (0.040 \times 0.21) - (-0.266 \times 0.08) - (0.109 \times 0.02) + (0.217 \times 0.37) + (0.137 \times 0.27) + (0.274 \times 0.37) + (0.109 \times 0.17) - (- 0.50 \times 0.09) - (-0.092 \times 0.16)$$

$$\mathbf{f}_{1sup} = 0.96.$$

The same computation (with the appropriate weights) will thus give the following values for the beer prototype for all the components:

	$\mathbf{f}_1$	$\mathbf{f}_2$	$\mathbf{f}_3$	$\mathbf{f}_4$	$\mathbf{f}_5$
Prototype	0.96	–0.24	–.43	0.10	0.06

As shown in Figure 3.13A, additional observations can be added to the graphs of the active observations using the component values as coordinates. The position of the beer prototype indicates that it differs from *BlueMoon* and appears as the average beer of the other beers.

Table 3.18. A supplementary observation. A new beer was evaluated by the authors with the same variables as the original 6 beers. The second line is obtained by subtracting the average of the original 6 beers (data in Table 1) from the values of the prototype beer.

	Bitter	Acid	Sweet	Astringent	Alcoholic	Hop	Cereal	Toasted	Yeast	Hay	Malt	Fermented
Prototype	4,2	2,4	1,5	3,4	4,1	4,6	3,9	3,4	3,6	3,3	3,8	3,3
Prototype – Mean	**0,42**	**–0,14**	**–0,73**	**0,21**	**–0,08**	**0,02**	**0,37**	**0,27**	**0,37**	**0,17**	**–0,09**	**–0,16**

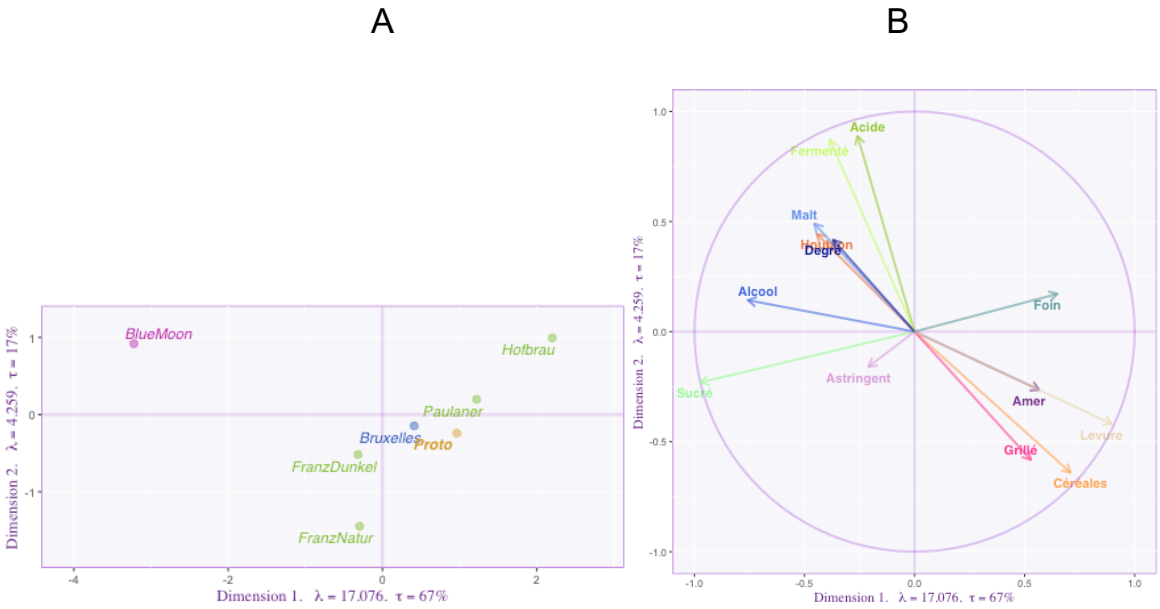

Figure 3.13. A) An additional beer (the beer "Prototype"), B) An additional variable: the alcohol degree.

To project supplementary variables, suffice to calculate their correlations with the factors of the analysis and to use these correlations as coordinates to place them in the circle of correlations (one can also re-normalize these coordinates to make them commensurate with the weights). Here we show how to integrate an additional variable into the correlation circle. For example, we measured the alcohol degree of the beers reported below:

	Bruxelles	FranzDunkel	FranzNatur	Hofbrau	Paulaner	BlueMoon
Degré d'alcool	4.5	5.0	5.0	5.1	5.2	5.4

The correlations between the alcohol degree and the first two components—reported below—are then used as coordinates to place the supplementary variable *alcohol degree* in the correlation circle (see Figure 3.13B).

	$\mathbf{f}_1$	$\mathbf{f}_2$	$\mathbf{f}_3$	$\mathbf{f}_4$	$\mathbf{f}_5$
Alcohol degree	–0.37	0.42	–0.75	0.27	0.24

As shown by the examination of these correlations and Figure 3.13B, the alcohol degree is close to the perceived alcohol intensity (but without being identical) but an important part of its variance is explained by the third dimension (unlike the perceived intensity which is explained mainly by the first dimension).

To sum up: Additional variables and observations are mainly used to facilitate the analysis's interpretation or to add a variable or an observation not considered in the original analysis.

3.3.6. Rotation

The interpretation of the results of a PCA is particularly easy: when the weights of the variables or the observations for a component are either very large or very small. When these conditions are met, we say that the structure of the PCA results is a *simple* structure. It is not always so, and to facilitate the interpretation one sometimes tries to "simplify" the results. To do so, the most popular approach looks for a rotation of the axes representing the components to obtain a simpler structure. Although there is quite a number of methods to perform these rotations (see Abdi, 2003, for some of them), most of the applications use an approach developed by Henry Kaiser (in his doctoral thesis in 1956 and published in 1958) sometimes (curiously), called “the little Jiffy” (Kaiser, 1970). This approach has two steps: the first—which we have already seen—is to identify the number of components that we want to retain (using, of course, preferably the Kaiser criterion), the second step performs a rotation of the components to obtain a simple configuration. In most cases, the method used to perform this rotation will be the procedure called “Varimax rotation” which seeks to maximize the sum of the fourth powers of the component weights (in contrast to the PCA which maximizes the square of the weights). For the example of the beers, Table 3.8 suggests that three components are needed to reconstruct the essential variance of the variables and we will perform the rotation in the spaces of the first three principal components. Figure 3.15 shows the effect of

the Varimax rotation applied illustrated with the biplot (note: although the biplot represents observations and variables, the rotation is calculated only from the variables). In this example, the rotation mainly affects the first dimension (the rotation in the plane of the first two components corresponds to an angle of about 20 degrees). After rotation, the first component now completely explains the variable *Sweet* and clearly shows the opposition between, on the one hand, *Blue Moon* (the sweetest) and, on the other hand, *Paulaner* and *Hofbrau* (the least sweet). As can be seen in Figure 3.15, the rotation can make the interpretation of the PCA easier and the question then is whether or when to apply a rotation to the PCA results. Here opinions diverge strongly and tend to align with national statistical schools. Thus, the Anglo-Saxon school (very close to psychometrics) systematically recommends the use of rotations with as a first step to keep only the components whose eigenvalues are higher than the average of the eigenvalues (i.e., the Kaiser criterion, see Figure 3.5; note that here this criterion—if applied strictly—would indicate that a rotation is unnecessary, because at least two components are needed to perform a rotation). Note that this first step is equivalent to estimating the number of real components in the data—a problem far from being solved. The European school (especially the French school) considers that rotations are unnecessary, in part because it is not possible to define these rotations as a specific criterion since the procedure requires first estimating the dimensionality of the data. In practice, it seems that the rotation is useful when the data indicate that a small number of components explain the essential variance and that the small components can be considered as "random noise." More recently, several procedures alternative to rotation, grouped under the name of "sparsification methods (cf. Guillemot et al., 2019)," have been developed to obtain simple structures. These methods still used too rarely should however become more popular in the years to come.

To sum up: A rotation procedure often facilitates the interpretation of the results when the data show a clear separation of components representing the signal and components representing the noise.

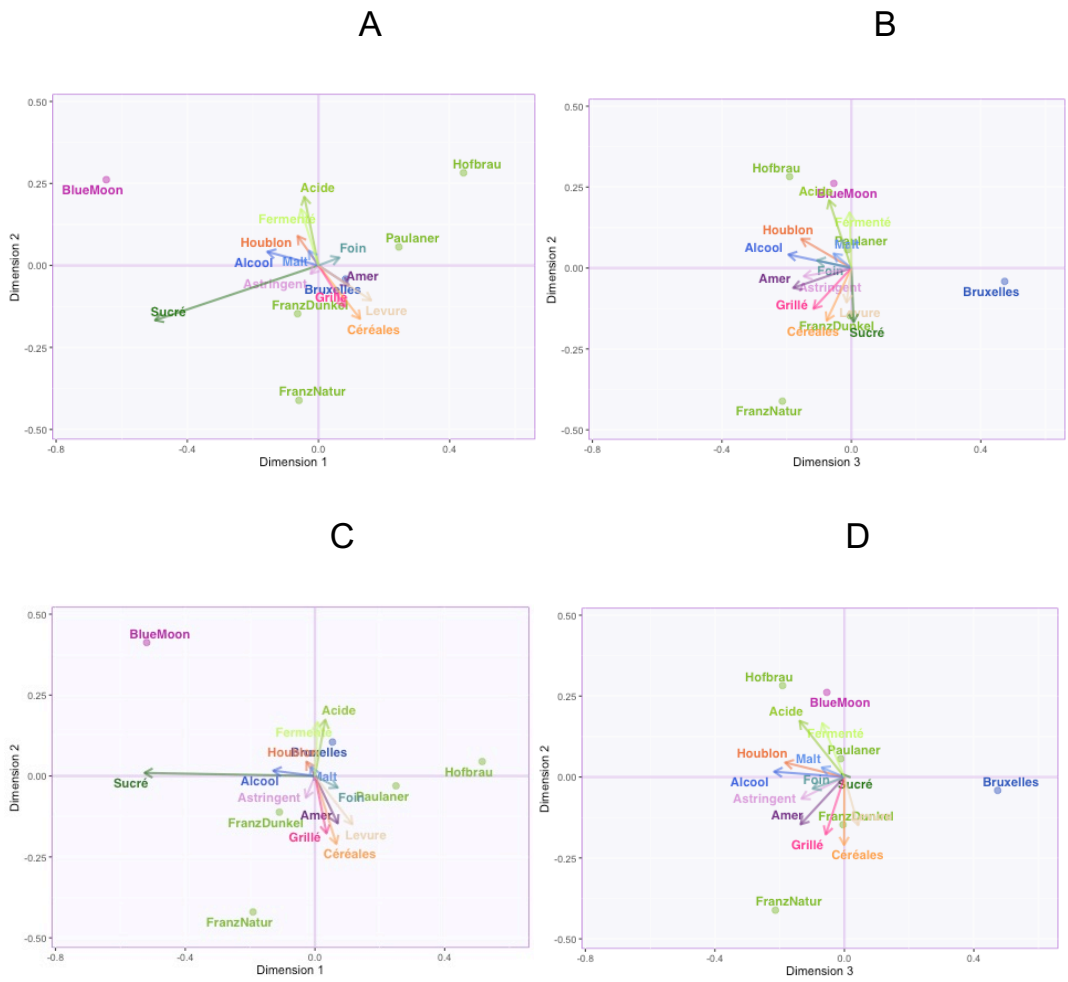

Figure 3.15. Effect of the Varimax Rotation. A) Biplot for principal Components 1 and 2 B) Biplot for principal Components 3 and 2. C) Biplot after rotation for principal Components 1 and 2 D) Biplot after rotation for principal Components 3 and 2. We see, for example, that after a (clockwise) of approximately 20 degrees (cf. A and C), the variable '*Sweet'* is now completely aligned with the first component.

4. Normed or Correlation PCA

In the previous example of a centered principal component analysis (a.k.a., a covariance PCA) all the variables measured the intensity of sensory attributes with the same scale going from 0 to 7. In this case, to differentiate between the products a variable with a large variance is more important than a variable with a small variance, and therefore the unit of measurement should be kept in order to keep the differences in variance. In a lot of cases, however, the variables used for the analysis are measured with units that are either incomparable or incommensurable (for example: a weight measured in grams and a temperature measured in degrees). In these cases, we need to normalize the variables to be able to integrate them in the same analysis (not normalizing would be like comparing apples and oranges). The normalized

variables are often called variables without unit and so normalized variables are directly comparable (because of being unitless, they have, paradoxically, the same unit). We use mostly two methods to normalize variables. The first method converts the variables to "*Z*-scores" (a favorite transform in the social sciences). With this transformation, the variance of a variable is equal to one and its sum of squares is equal to the number of observations (or the number of observations minus one, depending upon the variant of *Z*-scores). The other method (preferred by statisticians) creates normed variables. This procedure first centers the variable (i.e., subtract the mean of the variable from all the values of this variable) and then normalizes the variable by dividing all values of the variable by the square root of its sum of squares (a quantity called the *norm* of the variable hence the name "normed variables"). Normed variables have a sum of squares equal to 1. With the normed variables, the covariance between two variables becomes their correlation. We call a PCA performed on normed variables a normed or a correlation PCA. We illustrate this approach with our next example (on mint chewing gums).

4.1. A First Look at The Data

The data for the chewing gum example are given in Table 4.1. This table shows 16 products (chewing gums) evaluated by four variables (here similarly to the previous example, the scores are the average of 10 panelists). Different scales were used to evaluate these products: *Menthol* was measured with a scale going from 1 to 10, *Spearmint* and *Peppermint* were measured with a scale going from 0 to 5, and the last variable, *Long Lasting*, was evaluated with a scale going from 0 to 100.

Table 4.1. The Chewing Gum Tasting Data Set

Long Names	Short Names	Menthol	Spearmint	Peppermint	LongLasting
Hollywood Ice fresh	H.fresh	7	2	4	60
Oral B Hollywood verte	H.verte	5	3	4	60
Oral B Hollywood forte	H.forte	7	1	5	70
Hollywood Green fresh	H.green	6	4	5	40
Freedent white	F.white	7	1	4	45
Hollywood Power fresh	H.power	8	3	3	50
Mentos Pure fresh	M.fresh	7	2	4	65
Freedent verte	F.verte	4	5	1	55
Trident verte	T.verte	3	5	1	85
Trident blue	T.blue	4	2	4	90
Airwaves extreme	A.extreme	8	1	3	75
Arwaves verte	A.verte	5	4	2	65
Freshman's Friend original	S.ori	10	0	3	70
Bonsai methe forte	B.forte	9	1	4	55
Bonsai chlorophylle	B.chloro	6	4	4	55
Chicza mayan rainforest menthe sauvage	C.menthe	1	2	1	20

4.2. The Normalized Data

Since the variables from Table 7.1 are measured with different units they are not comparable and therefore the first step of the analysis centers and normalizes the variables. For example, we center the variable *Long-Lasting*, whose mean is equal to 60, by subtracting 60 from all its values. The new centered variable now take the values: 0, 0, 10, –20, … , 10, –5, –5, and –40. To normalize the variable, we now take each of the values of the centered variable, square them, and add them all:

$$0^2 + 0^2 + 10^2 + \cdots + 10^2 + (-5)^2 + (-5)^2 + (-40)^2 = 4,400$$

The square root of this sum is the norm of the variable *Long-lasting*: it is equal to $\sqrt{4{,}400} = 66.33$. The variable *Long-Lasting* is then normed (we can also say normalized) by dividing

the values of the centered the variable by its norm. The result of these operations is found in the column *Long-lasting* of Table 4.2.

Table 4.2. The Gum Tasting Data Set after Centering and Normalization

	Menthol	Spearmint	Peppermint	LongLasting
H.fresh	0.104	-0.083	0.144	0.000
H.verte	-0.118	0.083	0.144	0.000
H.forte	0.104	-0.250	0.337	0.151
H.green	-0.007	0.250	0.337	-0.302
F.white	0.104	-0.250	0.144	-0.226
H.power	0.215	0.083	-0.048	-0.151
M.fresh	0.104	-0.083	0.144	0.075
F.verte	-0.229	0.417	-0.433	-0.075
T.verte	-0.340	0.417	-0.433	0.377
T.blue	-0.229	-0.083	0.144	0.452
A.extreme	0.215	-0.250	-0.048	0.226
A.verte	-0.118	0.250	-0.241	0.075
S.ori	0.438	-0.417	-0.048	0.151
B.forte	0.327	-0.250	0.144	-0.075
B.chloro	-0.007	0.250	0.144	-0.075
C.menthe	-0.563	-0.083	-0.433	-0.603

Before proceeding to the analysis, a look at the table of correlation (see Table 4.1) gives a first idea of the correlation structure of the data. This table shows that *Menthol* and *Peppermint* are positively correlated (r = .55) and that they are both negatively correlated with *Spearmint* (with values of r equal to, respectively, –55 and –.42), whereas *Long Lasting* has very low correlations with all the other variables (almost null for some). This pattern indicates that the variables *Menthol, Peppermint*, and *Spearmint* capture the same sensory characteristic (with *Spearmint* being on one side of the scale versus *Menthol* and *Peppermint* on the other side of the scale), and so these variables should all contribute to the same component. *Long Lasting* being almost orthogonal to these three variables should determine a second dimension by itself. So, even though we have four variables, the pattern of

correlations suggests that two components are likely to extract most of the information in the data.

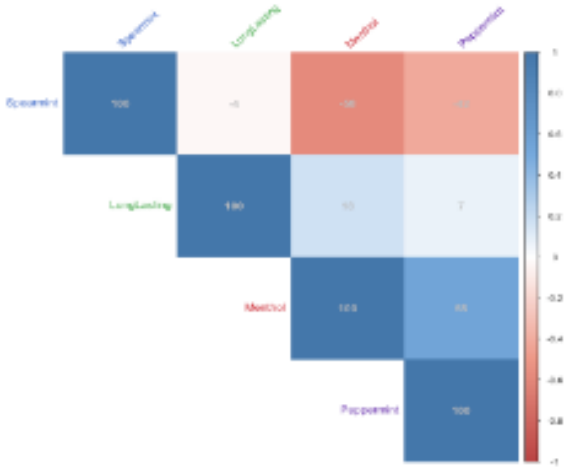

Figure 4.1 *Heatmap of the Correlations between the Variables of the Chewing Gum Data.*

4.3. How to interpret the results of a normed PCA

The steps to interpret a normed PCA are roughly the same as the steps used to interpret a covariance PCA (such as the previous Beer example).

First, we identify the number of relevant components. Second, we interpret the results from the point of view of the variables and identify the variables important for the components (using the variable weights), but we also identify the components important for the original variables (using the variable correlations with the components). Third, we interpret the analysis from the point of view of the observations and identify the observations important for the components (using the contributions), but also identify the components important for the observations (using the cosines of the observations with the components).

4.3.1. How many components: The scree Plot

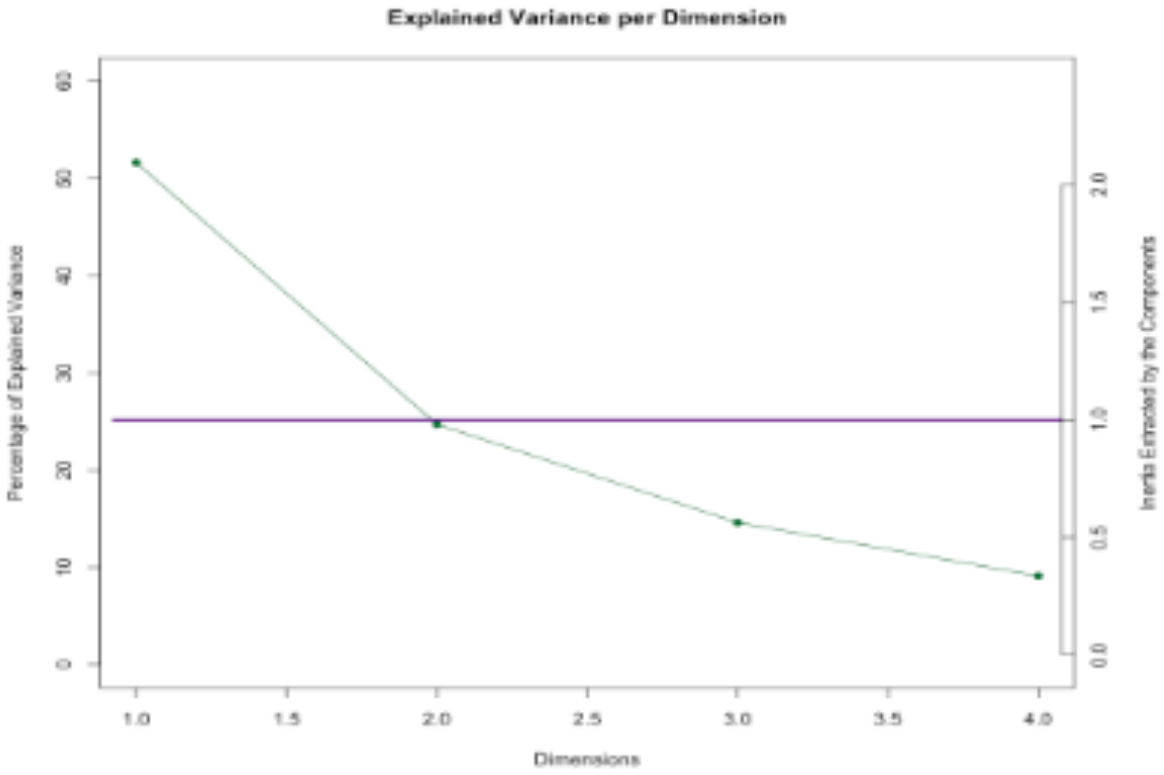

Figure 4.2 Chewing Gums. The scree plot of the eigenvalues. The Kaiser's line represents the average eigenvalue (and average percentage of variance): Components above the Kaiser's line are kept for the interpretation.

Here again the first step is to identify the number of components to keep for the analysis. And once again, we can use Kaiser's criterion (despite its shortcomings). For a normed PCA, this criterion becomes a simple rule because in a normed PCA all variables have a squared norm (i.e., a sum of squares) equal to 1 and therefore the total Inertia of the data is simply equal to the number of variables (i.e., here 4). This property leads to the well-known criterion (curiously called by Kaiser the "little Jiffy" criterion) for normed PCA: "Keep only the components with an eigenvalue larger than one." Here if we use Kaiser's criterion (see Figure 4.2, and Table 4.3) only the first component meets Kaiser's criterion, but the second component is very close to having an eigenvalue equal to 1 (i.e., $\lambda_2 = .998$), therefore we will keep the first two components for the rest of the analysis. Note that the elbow test would also keep two components.

4.3.2. Interpreting components, variables and observations

Table 4.3. The four principal components for the ten gums

	f_1	f_2	f_3	f_4
H.fresh	-0.188	-0.030	0.047	-0.008
H.verte	0.040	-0.002	0.172	-0.102
H.forte	-0.409	0.064	0.079	-0.179
H.green	0.007	-0.283	0.430	0.045
F.white	-0.245	-0.282	-0.051	-0.056
H.power	-0.036	-0.122	0.007	0.249
M.fresh	-0.200	0.044	0.044	-0.020
F.verte	0.618	0.049	-0.037	0.183
T.verte	0.615	0.488	-0.047	0.028
T.blue	-0.056	0.405	0.054	-0.339
A.extreme	-0.279	0.182	-0.224	0.031
A.verte	0.329	0.145	-0.014	0.101
S.ori	-0.496	0.081	-0.346	0.136
B.forte	-0.404	-0.131	-0.074	0.097
B.chloro	0.076	-0.042	0.275	0.078
C.menthe	0.628	-0.567	-0.314	-0.244
Sum of Squares (Eigenvalue)	*2.064*	*0.988*	*0.584*	*0.364*

Table 4.4. The variable weights for the principal components

	f_1	f_2	f_3	f_4
Menthol	-0.611	0.019	-0.076	0.788
Spearmint	0.553	0.191	0.648	0.487
Peppermint	-0.544	-0.109	0.757	-0.346
LongLasting	-0.158	0.975	-0.041	-0.150
Sum of Squares	*1.000*	*1.000*	*1.000*	*1.000*

The values in Tables 4.3, 4.4, and 4.5 are used to draw the maps in Figure 4.3. These tables and figure confirm the interpretation derived from the correlation matrix and heatmap: The information in the data corresponds to two dimensions for the variables (these are well represented on the circle of correlation). Some chewing gums are not well represented on the first two components (e.g., *H.verte* and *B.chloro*), however, because most of the chewing gums are well explained by the first two dimension, we can consider that these two dimensions extract the essential information in the data for both the variables and the observations but that these gums (i.e., *H.verte* and *B.chloro*) do not match the general pattern.

Table 4.5. The Variable Weights (Scaled to Eigenvalues) for the Principal Components

	f_1	f_2	f_3	f_4
Menthol	-0.878	0.019	-0.058	0.475
Spearmint	0.795	0.190	0.495	0.294
Peppermint	-0.781	-0.108	0.578	-0.209
LongLasting	-0.226	0.969	-0.031	-0.090
Sum of Squares (Eigenvalues)	*2.064*	*0.988*	*0.584*	*0.364*

We conclude, from the first component (that explains more than half of the inertia in the data), that this first component expresses a dimension that we could call *Mint* that confound *Menthol* and *Peppermint*, which are both opposed to the mint *Spearmint*. According to this interpretation the negative side of Component 1 would represent "high Mint" and the positive side "low Mint." This way, the chewing gums evaluated with high scores for Mint are also evaluated with high scores for *Peppermint* but low scores for *Spearmint* (and vice versa). Also, because the variable *Long Lasting* is roughly orthogonal to the first *Mint* component and is the sole contributor to the second dimension, we conclude that the lasting effect for the taste of a chewing gum is a property independent from the type of mints used in the formulation of these chewing gums.

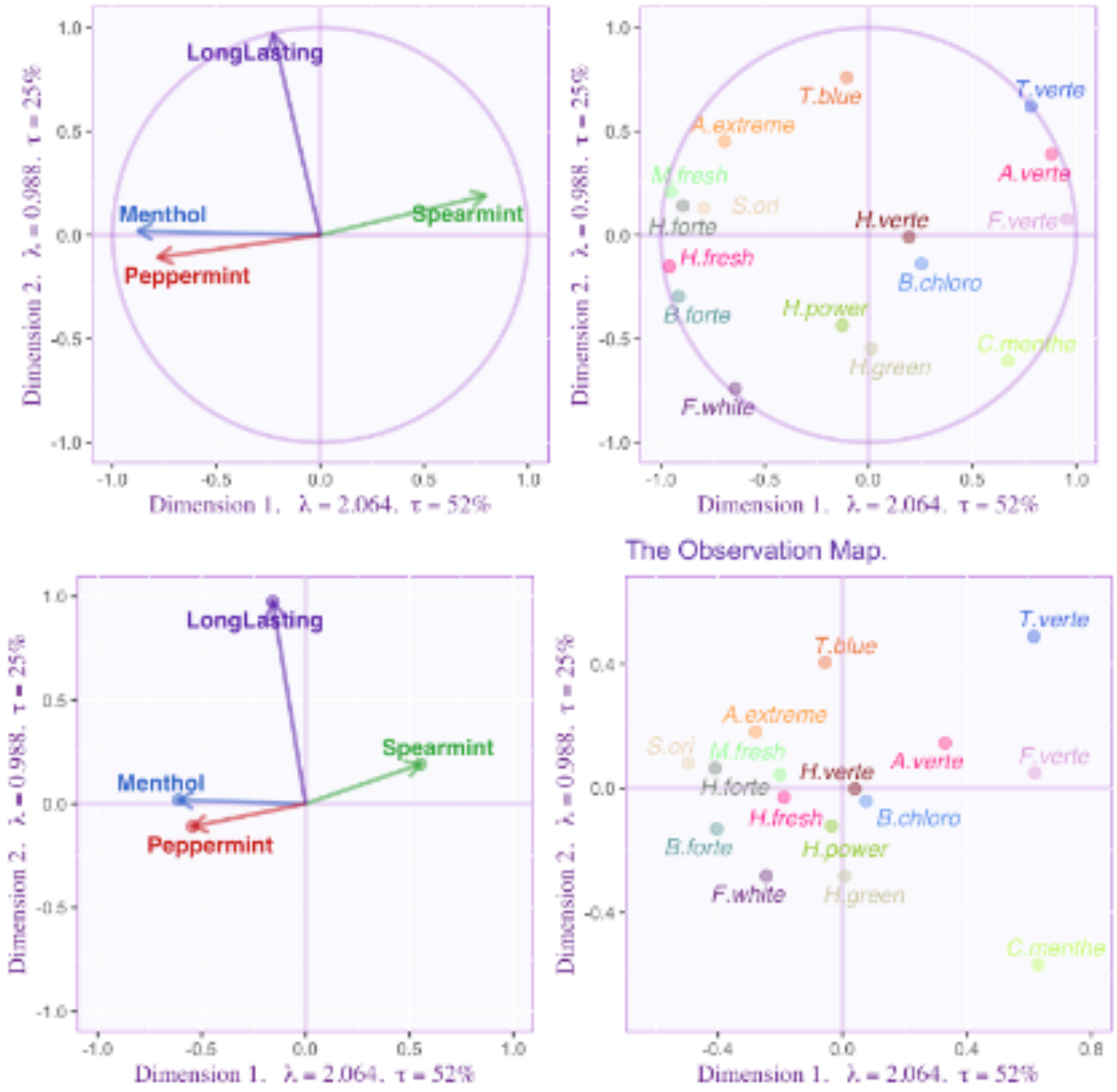

Figure 4.3 Chewing Gums. Components 1 and 2. Circles of Cosines (Descriptors and Gums). Descriptor weights and Gums Factor Scores

4.4. Biplot and Rotation

When the number of variables and (to a lesser extent) the number of products are relatively small, a simultaneous representation (i.e., a "biplot") is a convenient way to display the gist of the analysis with a single graphic (as shown in Figure 4.4A). This representation is even more readable after applying a Varimax rotation in the first two dimensions of the PCA (Figure 4.4B): now, the first component is simply the type of mint taste and the second component: the duration of the taste of gum in the mouth.

A B

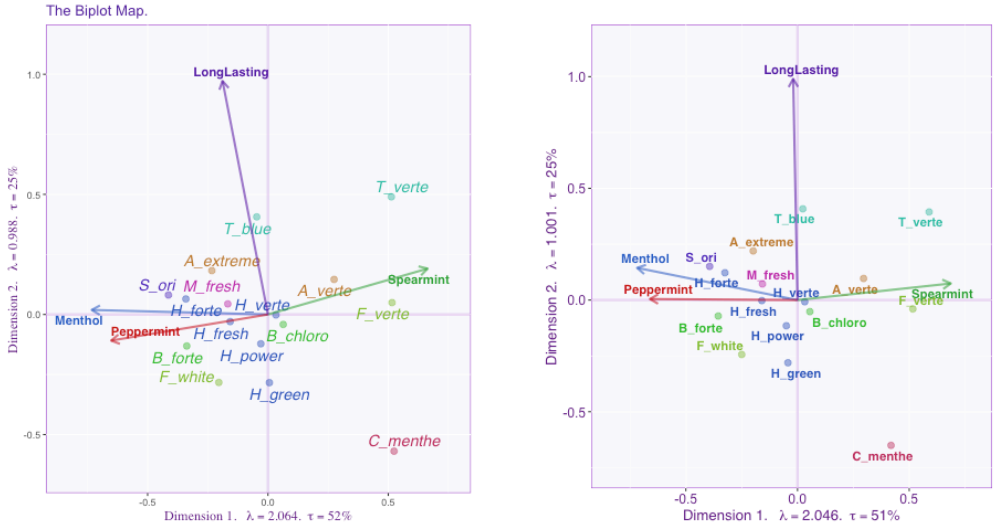

Figure 4.4. Chewing-gums: A) standard biplot B) Varimax rotated biplot.