7.3 Linear Unmixing

a For principal component analysis, we have written a model equation (7.1.i),

$$\boldsymbol{x}_c = \boldsymbol{S}\,\boldsymbol{C}^T + \boldsymbol{r} \ ,$$

whose structure makes it appropriate for many applications in which a number p of components **combine linearly** with respect to $m > p$ measured values.

For example, p chemical substances yield (optical or chromatographic) **spectra** $\underline{c}_k = [c_k^{(1)}, ..., c_k^{(m)}]$, which characterize them. The spectra measure intensities $c^{(j)}$ – absorption or reflection – for m waves lengths or periods. A graphical representation of 121 near infrared (NIR-) spectra, which are measured in the course of a chemical reaction, is found in the introduction (1.2.g). For a mixture without a chemical reactions, the spectra mostly combine at least approximately linearly (according to the Beer-Lambert law). If a chemical reaction occurs, the spectrum of the resulting new substance comes into the mix.

The spectrum of the ith mixture is thus, up to measurement error, $S_i^{(1)}\underline{c}_1 + S_i^{(2)}\underline{c}_2 + ... + S_i^{(p)}\underline{c}_p$, where the $S_i^{(k)}$ indicates the proportions of the substance k in the ith mixture. The measured spectra $\boldsymbol{X}$ thus follow the stated model with $\underline{\mu} = \underline{0}$. The spectra $\underline{c}_k$ of the substances form the columns of the matrix $\boldsymbol{C}$.

Since spectra and proportions can not be negative, it must be true that $S_i^{(k)} \geq 0$ and $C_j^{(k)} \geq 0$ for all i, j and k. We call the model with these constraints the **linear mixing model**.

b In the same way, there is overlaying of

- air pollutants that come from p sources, characterized by their “source profiles” $\underline{c}_k$. Their contribution $S_i^{(k)}$ to a pollution measurement $\underline{X}_i$ is dependent on their impact at the ith time point, which depends on their activity and the transportation from the source to the measurement point.
- chemical elements in rocks that are composed of multiple basic rocks,
- trace elements in well water that has passed through different rock strata.

c **Model.** We write the equation again, but now for uncentered random variables instead of for centered data:

$$\boldsymbol{X} = \boldsymbol{S}\,\boldsymbol{C}^T + \boldsymbol{E} \ .$$

Writing an error term $\boldsymbol{E}$ instead of a “residual term” $\boldsymbol{r}$ indicates that here, as usual, we assume normally distributed deviations, i.e. $\underline{E}_i \sim \mathcal{N}\langle \underline{0}, \boldsymbol{\Sigma} \rangle$, indepedent, for the deviations $\underline{E}_i$ of the ith observation, as in multivariate regression.

d If the source profiles (spectra) $\underline{c}_k$ are all known, then the contribution $S_i^{(k)}$ for each observation i can be determined separately with the help of a multiple regression estimation. In fact, $\underline{X}_i = \boldsymbol{C}\underline{s}_i + \underline{E}_i$, and this is the matrix form of a multiple regression model with the “X matrix” $\boldsymbol{C}$ and the coefficients $s_i^{(k)}$, $k = 1, 2, ..., p$. However, the deviations $E_i^{(j)}$ hardly have the same variances. For a good definition, we need assumptions about these variances and then weighted regression (* or non-linear for a transformed version of the variable of interest, for example $\log\left\langle X_i^{(j)} \right\rangle = \log\left\langle \underline{C}_j^T \underline{s}_i \right\rangle + \widetilde{E}_i^{(j)}$).

e **Linear Unmixing.** More interesting is the case in which both the **source profiles** $\underline{c}_k$ as well as the contributions $\underline{S}_i$ must be **estimated** from the data. We are then talking about **linear unmixing**. With a combination of statistical methods, application specific particularities, and expertise, this can often be achieved well.

The main difficulty lies in the fact that $\boldsymbol{S}$ and $\boldsymbol{C}$ in the model are not unique – so **not identifiable**: For every invertible matrix $\boldsymbol{T}$, $\boldsymbol{S}$ and $\boldsymbol{C}$ can be replaced with $\widetilde{\boldsymbol{S}} = \boldsymbol{S}\boldsymbol{T}$ and $\widetilde{\boldsymbol{C}} = \boldsymbol{C}(\boldsymbol{T}^T)^{-1}$ without, for equal errors $\boldsymbol{E}$, changing the data $\boldsymbol{X}$.

f **Estimation of the Subspace.** By statistical means, we can therefore initially estimate only the “error-corrected observations” $\widetilde{\boldsymbol{X}} = \boldsymbol{S}\boldsymbol{C}^T = \widetilde{\boldsymbol{S}}\widetilde{\boldsymbol{C}}^T$.

If we assume independent $E_i^{(j)} \sim \mathcal{N}\langle 0, \sigma^2 \rangle$ with equal variances, a version of **principal component analysis**, for which the data is not centered, gives the best estimation of $\widetilde{\boldsymbol{X}}$. (S: `prcomp(..., center=FALSE)`.) If we believe we know the relationship between the variances $\sigma_j^2 = \mathrm{var}\left\langle E_i^{(j)} \right\rangle$, we can divide each of the variables $X^{(j)}$ by σ_j and then apply the non-centered principal component analysis. If we know nothing about the variances, factor analysis gives an estimation (see 7.4.b).

g The error-corrected observations lie in a space with dimension p. In this **subspace**, the “**axes**” are initially **undetermined**. In application, various considerations lead to a reasonable determination of the axes. A harmless uncertainty stems from the fact that each source profile, so each column of $\boldsymbol{C}$, can be multiplied by a scalar factor and the corresponding column of $\boldsymbol{S}$ be divided by the same number and we still get the same error-corrected observations. In this sense, the source profiles must therefore be appropriately standardized to render the model unique.

h ▷ In the **NIR spectra example**, in the representation of the first four principal components (7.1.j) we can recognize four phases of the reactions, in each of which the process moves in one direction. We can use the corresponding four direction vectors as new axes and calculate the corresponding coordinates or scores. If we plot them against the temporal order of the observations, we get the representation 1.2.g (ii), shown in the introduction. ◁

i If we want to find appropriate axes in other examples of linear unmixing, then the inequalities $S_i^{(k)} \geq 0$ and $c_j^{(k)} \geq 0$ and existing subject-based knowledge about the source profile (spectra of the chemical substances) $\underline{c}_k$, i.e. the columns of the matrix $\boldsymbol{C}$, are helpful. We will briefly comment on both.

j **Non-Negativity.** The non-negativity of the $S_i^{(k)}$ and $c_j^{(k)}$ already significantly limits the freedom of choice. The linear mixing model says that each observation vector $\underline{X}_i$ is a linear combination of the source profiles $\underline{c}_k$, specifically one with non-negative coefficients, a so-called *convex* linear combination.

We illustrate this concept in the three dimensional space, so for $m = 3$ variables and $p = 2$ sources (Fig. 7.3.j). The two source profiles $\underline{c}_1$ and $\underline{c}_2$ lie in a plane. The possible mixtures $\underline{X}_i$ lie, up to the deviations $\underline{E}_i$, in this same plane — but not only that, they also lie in the sector between the two source profile vectors.

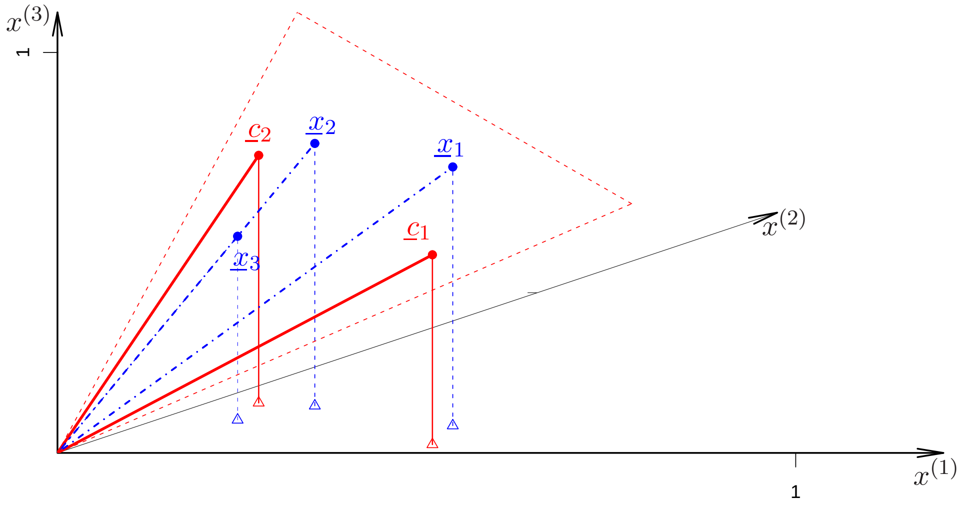

Figure 7.3.j: Illustration of the linear mixing model for 2 source profiles $\underline{c}_1$ and $\underline{c}_2$

If we now assume that, for certain observations, only a single source is active, then these observation vectors are recognizable as the most extreme vectors of all those observed. For two sources, they form the edges of the sector, for three sources the edges of a triangular pyramid with peak at the null point, etc.

This property can be used to determine the source profiles graphically or with a numerical algorithm that determines the edges of a minimal (hyper) pyramid that envelops the observations.

k **Compositional Data.** In certain applications, the data consist of proportions, for example of basic rock types in blocks of rock. The variables then add up to 100% or another fixed number.

If this is not fulfilled by the data, it can be reasonable to standardize in this sense, so dividing each entry $X_i^{(j)}$ by the row sum $\sum_j X_i^{(j)}$ (and to multiply by 100%). The source profiles, as mentioned above, have to be standardized somehow, which will also be done in this way.

Because of the standardization, such data already lie in a subspace. Fig. 7.3.k shows how this looks in the case of 2 sources.

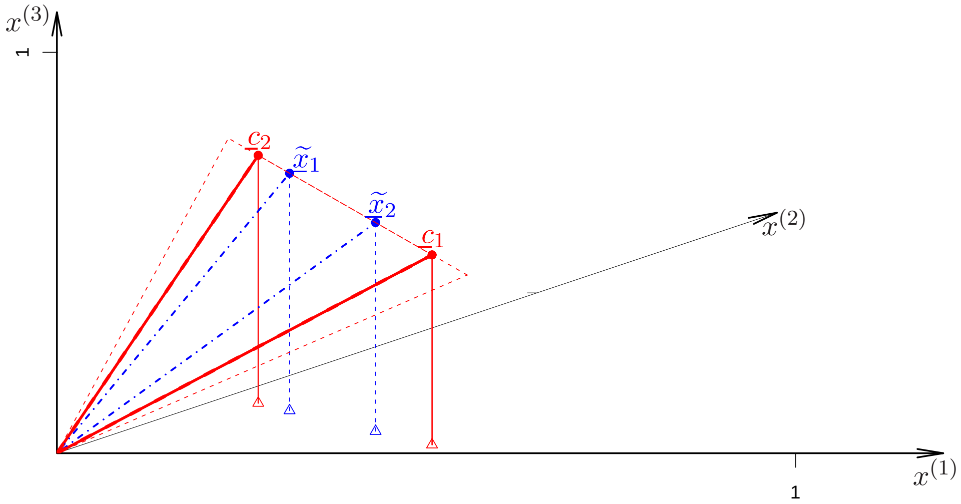

Figure 7.3.k: Illustration of the model for standardized data

Through standardization, the space, in which we must seek appropriate axes, becomes one dimension smaller and we find it with the usual principal component analysis that works with centered data.

l **Chemical Reactions.** For chemical reactions, the mixing prportions of the input substances and their spectra are usually known for the starting time point. Through the reaction, one or more new substances are created, while the substances of the initial mixture disappear in a clear, simple (stoichiometric) relationship. This helps to find the unknown spectra of the created substances. Note, however, that the conservation of elements introduces a linear constraint, and the natural directions in the subspace are given by weighted differences of spectra.

L **Literature:** Further explanation of the methodology is found in the literature under the name *linear mixing model* and *mass balance*; an appropriate starting point is Renner (1993). Basics are also found in books about chemometrics.

7.4 Factor Analysis

a The model 7.1.i has long been well known in psychology by the name **factor analysis**. In the typical application, $X_i^{(j)}$ is the score that test subject i achieved for the jth test task. It is perceived as the result of a superposition $S_i^{(1)}C_j^{(1)} + S_i^{(2)}C_j^{(2)} \;(+ \ldots + S_i^{(p)}C_j^{(p)})$ of factors that are often interpreted as mathematical intelligence $S_i^{(1)}$ and verbal intelligence $S_i^{(2)}$ (and possibly other "dimensions" of intelligence), up to a random deviation $E_i^{(j)}$.

The **"factors"** $S^{(k)}$ are, however, not observable; they can only be discovered via that observed values $X_i^{(j)}$. Such random variables are called **latent variables** .

b **Model.** The model again has, as results from these considerations, the form $\boldsymbol{X} = \boldsymbol{S}\,\boldsymbol{C}^T + \boldsymbol{E}$, where usually the data $\boldsymbol{X}$ are assumed to be centered, in contrast to the linear mixture model. Factor analysis differs from principal component analysis in two respects:

- For the deviations $\underline{E}_i$ a normal distribution is assumed, $\underline{E}_i \sim \mathcal{N}_m\langle \underline{0}, \boldsymbol{\Sigma} \rangle$. Specifically, it is assumed that the deviations for the *variables* j, in other words the $E^{(j)}$, are independent of each other (not only the observations). The covariance matrix $\boldsymbol{\Sigma}$ is therefore assumed to be diagonal, $\boldsymbol{\Sigma} = \mathrm{diag}\langle \sigma_1^2, \sigma_2^2, ..., \sigma_m^2 \rangle$. This means that the dependencies between the variables $X^{(j)}$ are fully explained by the term $\boldsymbol{S}\,\boldsymbol{C}^T$. The factors k with the “scores” $\underline{S}^{(k)}$ are therefore more precisely called “common factors”, while the independent $E^{(j)}$ are labelled “unique factors”, and their variances σ_j^2, “uniquenesses”.
- The scores $S_i^{(k)}$ are also interpreted as random vectors and also modeled with a normal distribution $\mathcal{N}_p\langle \underline{0}, \boldsymbol{V} \rangle$.

This makes the $\underline{X}_i$ normally distributed observation vectors with

$$\underline{X}_i \sim \mathcal{N}_m\langle \underline{0},\ \boldsymbol{C}\,\boldsymbol{V}\,\boldsymbol{C}^T + \mathrm{diag}\langle \sigma_1^2, ..., \sigma_m^2 \rangle \rangle \ .$$

This is a probability model for which we can estimate the parameters σ_j^2 and the elements of $\boldsymbol{C}\,\boldsymbol{V}\,\boldsymbol{C}^T$. However, the splitting of $\boldsymbol{C}\,\boldsymbol{V}\,\boldsymbol{C}^T$ into the matrix $\boldsymbol{C}$ and the covariance matrix $\boldsymbol{V}$ of the scores are still not unique.

c **Number of Factors.** Unlike principal component analysis, where the number of factors that will be used for interpretation or representation of the data is fixed arbitrarily, in the factor analysis model this number is given by the above-mentioned requirement that the remaining “unique factors” $E^{(j)}$ are independent.

According to the general principle of the likelihood ratio test, we can check whether p factors are sufficient. This is the case if the likelihood for the factor model is only slightly smaller than the likelihood for a general multivariate normal distribution. This test is included in popular programs for fitting factor analysis models.

In application, it can also be helpful to use a model that, in this sense, has significantly too few factors, primarily if the number of observations is so large that even insignificant remaining correlations of the $E^{(j)}$ lead to significance of an additional factor.

d **Principal Factors.** As mentioned, in factor analysis there is also the problem that the factors themselves, i.e. $\boldsymbol{C}$, are actually indeterminable (see 7.3.e). To make them unique, there are various criteria that have been proposed.

A more theoretically motivated way to achieve uniqueness is to use as factors the principal components of the “observations in the subspace” $\widetilde{\underline{X}}_i = \underline{X}_i - \underline{E}_i$. This way we achieve that the scores $\underline{S}_i$ are independent and the matrix $\boldsymbol{C}$ is “pseudo-orthogonal” , $\boldsymbol{C}^T\boldsymbol{C} = \boldsymbol{I}$. This variant of factor analysis is called **principal factor analysis**.

e **Interpretability of the Factors.** This solution serves as a starting point for the determination of further possible designations of the factors. The goal of such a designation is that the factors should be easily interpretable. For this, the following can be useful:

- The scores should be standardized. They already have mean zero because the data is centered before the factor analysis. Variance 1 means that, for example, a score of 1 means a value one standard deviation higher than the mean and, according to the normal distribution, we can expect that only 1/6 of the population will fall higher than this. – This requirement is easy to fulfill by standardizing factors which have been determined in any suitable way at the end.
- The scores should be independent, so that each factor measures an aspect of the data independent from the other factors. However, it is probably wrong to seek independent factors that measure mathematical and verbal intelligence.
- A factor is often easily interpretable if it has a high correlation with only a few original variables $X^{(j)}$ and if these variables have low correlations with the other factors.

The correlations are known as "**loadings**"

$$\lambda_j^{(k)} = \text{corr}\left\langle X^{(j)}, S^{(k)} \right\rangle \ .$$

We have $\text{cov}\langle \underline{X}, \underline{S} \rangle = \text{cov}\langle \boldsymbol{C}\underline{S} + \underline{E}, \underline{S} \rangle = \boldsymbol{C}\ \text{var}\langle \underline{S}, \underline{S} \rangle$. If the original variables are standardized in the univariate sense and the factors are multivariately standardized, then the matrix $\boldsymbol{C}$ reflects the loadings. Careful! For the principal factor analysis (without standardization of the factors) this is not true!

f **Rotations.** The standardization of the scores of the principal factor analysis means that their covariance matrix is the identity matrix, since they were already uncorrelated. The standardized principal factors are therefore also standardized in the multivariate sense. Now we can transform these standardized factors with every orthogonal matrix and retain the multivariately standardized scores. The various proposals for such transformations with the goal of improving the interpretability are therefore known as **orthogonal rotations**. This would really be a redundancy if those (linear) transformations that do not observe orthogonality were not known traditionally as "**oblique rotations**"(!)

- The requirement that only a few loadings $\lambda_j^{(k)}$ should be large can be quantified via the variance of the $(\lambda_j^{(k)})^2$. This criterion is known as **varimax** and is usually used for optimization.
- If "oblique rotations" are allowed, the criterion **oblimin** is usually optimized, see literature.

g ▷ **In the voting example** we can initially determine the number of factors with the likelihood ratio test. In this case the answer is clear: the p-value for 2 factors is 0, but for 3 factors is 0.54. So, three factors are needed. Table 7.4.g shows the loadings for the varimax rotation. The first factor receives high loadings for the votes i, c and b, the second for f and e, the third for k, l and m. The "uniquenesses" for these variables tend to be smaller than for the rest. The votes n, h, and g show especially high uniqueness.

	Factor1	Factor2	Factor3	Uniqueness
a	0.518	0.576	-0.478	0.463
b	0.786	0.283	0.348	0.400
c	0.852	0.309	0.172	0.418
d	0.536	0.510	0.175	0.652
e	0.195	0.871	-0.216	0.709
f	0.346	0.922		0.626
g		0.635	-0.401	0.800
h	-0.751	0.555	-0.189	0.864
i	0.964			0.631
j	0.583		0.440	0.689
k	0.129		0.973	0.251
l	0.252	-0.123	0.957	0.071
m	0.252	-0.140	0.945	0.159
n	-0.131	0.721		0.963

Table 7.4.g: Loadings and uniqueness in the voting example with 3 factors and varimax rotation

The results are viewable graphically in Fig. 7.4.g. The contrast between French and German Switzerland is not shown as well as in the principal components representation and the urban-rural difference is even worse. In this example, factor analysis thus proves to be a poor choice. This is not surprising, since a factor structure is not especially plausible for the voting patterns. ◁

h **Significance of Factor Analysis.** Factor analysis should then be applied if the variables are chosen so that they reflect certain suspected factors – latent variables that can not be measured directly. If clear theoretical ideas are available about the relationships between observed and theoretically supposed latent variables, then we can directly build these assignments into a model. This leads to the so-called **structural equation models**.

L **Literature:** An introduction to this methodology is found in chapter 12 in Bortz (1977). Famous books are Harman (1960, 1967) and Lawley and Maxwell (1971). All very old books! I am still searching for newer literature.

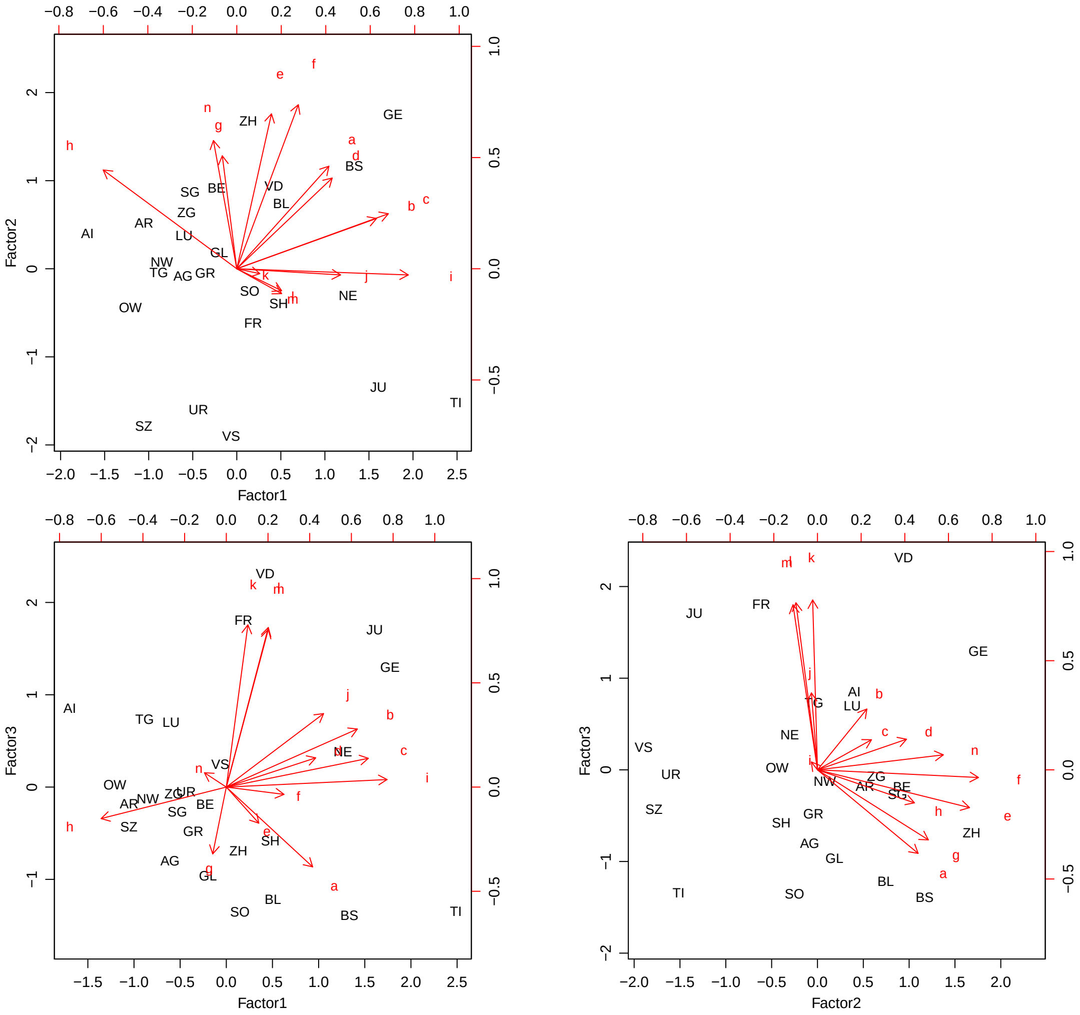

Figure 7.4.g: Biplot of the factor analysis in the voting example