5 Multivariate Analysis of Spectra

5.1 Introduction

a "Spectrum" means here: We measure the "intensity" for certain "wave lengths". Such a function characterizes a chemical mixture (or as a special case a pure substance).

There are many spectra in chemistry. For some of them, pure substances have a spectrum that consists of a single "peak". As long as the peaks are not overlapping, we can identify the different components of a mixture and their proportions.

b **NIR-Spectra** (near infrared): The NIR-Spectra of pure substances is "any" function with some more or less characteristic peaks. Hence, it's rather difficult to identify the type and the quantity of the different components based on the spectrum of a chemical mixture. On the other side, these spectra are very cheap: No extra processing is needed, they can be measured on-line.

Example **c** **Quality Control via NIR-Spectra** We have data of reflections of NIR-waves on 52 granulate samples with wave length 1100, 1102, 1104, ..., 2500 nm. Figure 5.1.c shows the spectra in "centered" form; for each wave length j the median value $\text{med}_i(X_i^{(j)})$ was subtracted from the $X_i^{(j)}$'s.

Wl.	1800	1810	...	2500
a	0.003097	0.017238	...	−0.02950
b	0.002797	0.016994		−0.03095
c	0.002212	0.015757		−0.03095
	...			
Z	0.001165	0.014237	...	−0.03110

Table 5.1.c: Data for the example "NIR-spectra" (for wavelengths larger than 1800nm).

Questions

- There are outliers. Are there other "structures"?
- The amount of an active ingredient was determined with a chemical analysis. Can we estimate it sufficiently accurate with the spectrum?

d In other applications we measure spectra to follow a reaction on-line. It is used for

- estimating the order of a reaction and to determine potential intermediate products and reaction constants,
- determining the end of a process,
- monitoring a process.

We can also automatically monitor slow processes of all kinds. For example stock-keeping: Are there any (unwanted) aging effects?

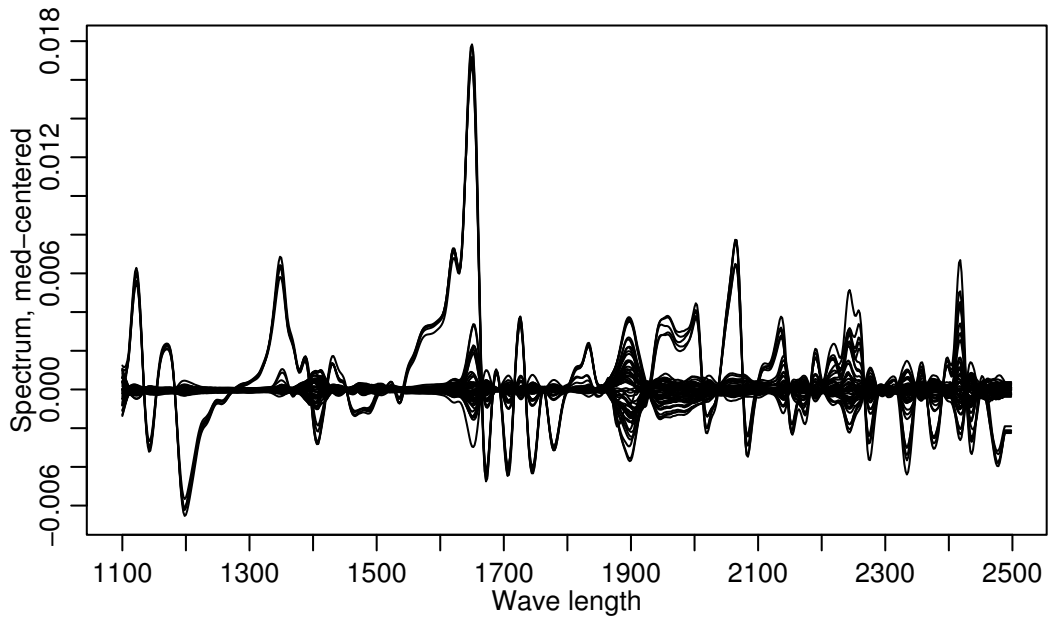

Figure 5.1.c: NIR-Spectra of granulate samples, centered at the median curve.

e For each observation (sample) we have many variables (a whole spectrum).

Questions

- Is it reasonable to plot the different samples on a plane or is it possible to catch most information and see structure from just a few dimensions (instead of using all variables)?
- Can we identify dimensions (with technical interpretation) in the high-dimensional space that contain most of the information?
- Is it possible to identify and to quantify the different components of a chemical mixture based on its spectrum?
- For a regression analysis, 70 variables (or 700 at a higher resolution) are too much if we only have 52 observations. How should we reduce dimensionality?

5.2 Multivariate Statistics: Basics

a **Notation** The vector $\underline{X}_i = [X_i^{(1)}, X_i^{(2)}, ..., X_i^{(m)}]^T$ denotes the ith spectrum. It's a point in m-dimensional space. Hence, for each observation we measure m different quantities.

Remark In statistics and probability theory vectors are usually column vectors. Row vectors are denoted by the symbol ^{T} (transposed vector).

This is inconvenient in statistics because the **data matrix**

$$\boldsymbol{X} = [X_i^{(j)}] \,,$$

that consists of n observations of m variables is built up the other way round: The ith row contains the values for the ith observation. For most applications this is a useful table (see e.g. the design matrix of a linear regression model). Here, it's often the other way round: In a table of spectra, a column often contains a single spectrum (i.e., it's one observation of a spectrum).

b **Definitions** We define the following quantities for an m-dimensional random vector $\underline{X} = [X^{(1)}, X^{(2)}, ..., X^{(m)}]^T \in \mathbb{R}^m$.

- **Expectation** $\underline{\mu} \in \mathbb{R}^m$
  $\underline{\mu} = (\mu_1, \ldots, \mu_m)^T$, where $\mu_k = \mathrm{E}[X^{(k)}]$, $k = 1, \ldots, m$.
  In other words: a vector that consists of the (univariate) expectations.
  We write $\underline{\mu}_X$ in situations where we also have other random variables.
- **Covariance Matrix** $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times m}$
  $\boldsymbol{\Sigma}$ is an $m \times m$ matrix with elements

$$\boldsymbol{\Sigma}_{jk} = \mathrm{Cov}(X^{(j)}, X^{(k)}) = \mathrm{E}\left[(X^{(j)} - \mu_j)(X^{(k)} - \mu_k)\right].$$

  We also use the notation $\mathrm{Var}(\underline{X})$ or $\mathrm{Cov}(\underline{X})$.
  Note that
  - $\boldsymbol{\Sigma}_{jj} = \mathrm{Cov}(X^{(j)}, X^{(j)}) = \mathrm{Var}(X^{(j)})$.
    This means that the diagonal elements of the matrix are the variances.
  - $\mathrm{Corr}(X^{(j)}, X^{(k)}) = \dfrac{\boldsymbol{\Sigma}_{jk}}{\sqrt{\boldsymbol{\Sigma}_{jj}\boldsymbol{\Sigma}_{kk}}}$.

  Again, sometimes we write $\boldsymbol{\Sigma}_X$ if we want to point out that this is the covariance matrix that corresponds to $\underline{X}$.

c **Linear Transformations**

- For a simple (one-dimensional) random variable: $Y = a + bX$, where $a, b \in \mathbb{R}$.
  Expectation: $\mathrm{E}[Y] = a + b\mu_X$.
  Variance: $\mathrm{Var}(Y) = b^2\sigma_X^2$.
- For random vectors: $\underline{Y} = \underline{a} + \boldsymbol{B}\underline{X}$, where $\underline{a} \in \mathbb{R}^m$, $b \in \mathbb{R}^{m \times m}$.
  Expectation: $\mathrm{E}[\underline{Y}] = \underline{a} + \boldsymbol{B}\underline{\mu}_X$.
  Covariance: $\mathrm{Cov}(\underline{Y}) = \boldsymbol{B}\boldsymbol{\Sigma}_X\boldsymbol{B}^T$.

d **Remark** The multivariate normal distribution $\underline{X} \sim \mathcal{N}(\underline{\mu}, \boldsymbol{\Sigma})$ is fully characterized by the mean $\underline{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. It is the most common distribution in multivariate statistics. See e.g. Chapter 15.3 in Stahel (2000).

Figure 5.2.d illustrates two two-dimensional normal distributions with the "contours" of their densities. The mean vector is responsible for the location of the distribution and the covariance matrix for the shape of the contours.

e **Estimators**

$$\begin{aligned}
\widehat{\underline{\mu}} &= \left[\overline{X}^{(1)}, \overline{X}^{(2)}, \ldots, \overline{X}^{(m)}\right]^T = \text{ vector of means} \\
\widehat{\boldsymbol{\Sigma}} &= \frac{1}{n-1}\sum_{i=1}^{n}(\underline{X}_i - \widehat{\underline{\mu}})(\underline{X}_i - \widehat{\underline{\mu}})^T \\
&= \text{matrix of the empirical variances and covariances.}
\end{aligned}$$

This means that

$$\widehat{\boldsymbol{\Sigma}}_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^{(j)} - \overline{X}^{(j)})(X_i^{(k)} - \overline{X}^{(k)}).$$

The covariance matrix plays a crucial role in multivariate models that are based on the **normal distribution** or that want to model **linear relationships**.

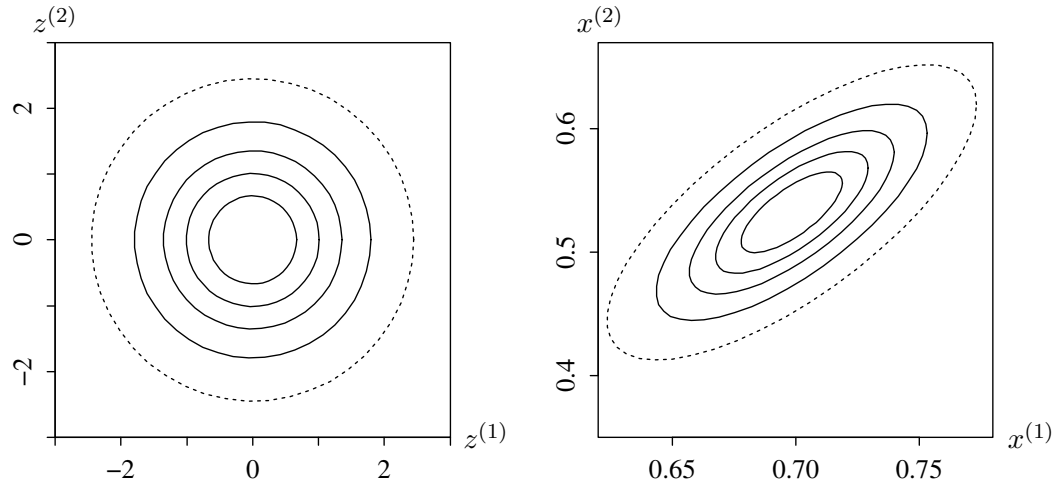

Figure 5.2.d: Contours of the probability densities for a standard normal (left) and a general (right) multivariate normal distribution.

5.3 Principal Component Analysis (PCA)

a Our goal is **dimensionality reduction**. We are looking for a few dimensions in the m-dimensional space that can explain "most of the variation in the data".

We define variation in the data as the sum of the individual m variances

$$\sum_{j=1}^{m} \widehat{\mathrm{Var}}(X^{(j)}) = \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j=1}^{m} (\widetilde{X}_i^{(j)})^2,$$

where $\widetilde{X}_i^{(j)}$ are the centered observations: $\widetilde{X}_i^{(j)} = X_i^{(j)} - \overline{X}^{(j)}$.

We want to find a new "coordinate system" with certain properties. This will lead to

- new basis vectors $\underline{b}_k$ ($\|\underline{b}_k\| = 1$), the so called **principal components**. The individual components of these basis vectors are called **loadings**.
- new coordinates $Z_i^{(k)} = \underline{\widetilde{X}}_i^T \underline{b}_k$, the so called **scores** (projections of the data on the directions above).

What properties should the new coordinate system have?

- The first basis vector $\underline{b}_1$ should be chosen such that $\mathrm{Var}(Z^{(1)})$ is maximal.
- The second basis vector $\underline{b}_2$ should be orthogonal to the first one ($\underline{b}_2^T \underline{b}_1 = 0$) such that $\mathrm{Var}(Z^{(2)})$ is maximized.
- And so on...

Figure 5.3.a illustrates the idea using a two-dimensional distribution.

To summarize, we are performing a **transformation to new variables**

$$\underline{Z}_i = \widehat{\boldsymbol{B}}^T (\underline{X}_i - \widehat{\underline{\mu}}),$$

where the transformation matrix $\widehat{\boldsymbol{B}}$ is orthogonal.

It can be shown that $\widehat{\boldsymbol{B}}$ is the matrix of (standardized) eigenvectors and $\widehat{\lambda}_k$ are the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_X$.

Remember that $\widehat{\boldsymbol{\Sigma}}_X$ is a symmetric matrix and therefore we can decompose it into

$$\widehat{\boldsymbol{\Sigma}}_X = \widehat{\boldsymbol{B}}\, \widehat{\boldsymbol{D}}\, \widehat{\boldsymbol{B}}^T,$$

where $\widehat{\boldsymbol{B}}$ is the matrix with the eigenvectors in the different columns and $\widehat{\boldsymbol{D}}$ is the diagonal matrix with the eigenvalues on the diagonal (this is a fact from linear algebra).

Therefore we have

$$\widehat{\mathrm{Var}}(\underline{Z}) = \widehat{\boldsymbol{B}}^T \widehat{\boldsymbol{\Sigma}}_X \widehat{\boldsymbol{B}} = \widehat{\boldsymbol{D}} = \begin{bmatrix} \widehat{\lambda}_1 & 0 & \dots & 0 \\ 0 & \widehat{\lambda}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & \widehat{\lambda}_m \end{bmatrix}$$

$$\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_m \geq 0.$$

Hence, the individual components of $\underline{Z}$ are **uncorrelated** and the first component of $\underline{Z}$ has largest variance. By construction it holds that $\widehat{\lambda}_1 = \widehat{\mathrm{Var}}(Z^{(1)})$. It is the maximal variance of a projection:

$$\widehat{\lambda}_1 = \max_{\underline{b}:\|\underline{b}\|=1} (\widehat{\mathrm{Var}}(\boldsymbol{X}\underline{b})).$$

Accordingly for $\widehat{\lambda}_m$: It's the smallest variance.

Because the $\widehat{\lambda}_k$ are the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_X$, we know from linear algebra that

$$\sum_{k=1}^{m} \widehat{\lambda}_k = \sum_{k=1}^{m} \widehat{\boldsymbol{\Sigma}}_{kk} = \sum_{k=1}^{m} \widehat{\mathrm{Var}}(X^{(j)}).$$

Hence

$$\frac{\sum_{j=1}^{k} \widehat{\lambda}_j}{\sum_{j=1}^{m} \widehat{\lambda}_j}$$

is the **proportion of the total variance** that is explained by the first k principal components.

Of course we can always go back to the original data using the new variables by doing a simple back-transformation

$$\underline{X}_i - \widehat{\underline{\mu}} = \left(\widehat{\boldsymbol{B}}^T\right)^{-1} \underline{Z}_i = \widehat{\boldsymbol{B}}\underline{Z}_i = \sum_{k=1}^{m} Z_i^{(k)} \underline{b}^{(k)}.$$

b **Graphical Representation** By reducing dimensionality it gets easier to visualize the data. For that reason we only consider the first two (or three) components and forget about the other ones. Figure 5.3.b (i) illustrates the first two components for the "NIR-spectra" example (for technical reasons we only consider wave lengths larger than 1800 nm). We can see 5 outliers – they were already visible in the spectra. Figure 5.3.b (ii) shows the first three components of a principal component analysis without the outliers.

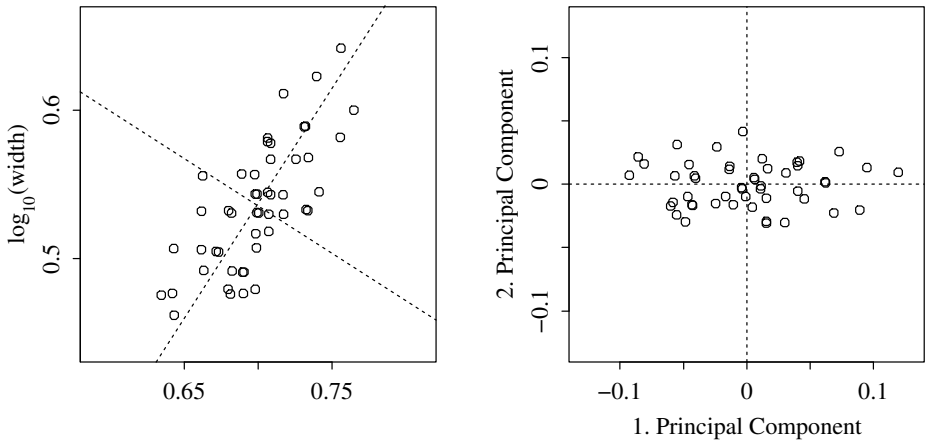

Figure 5.3.a: Principal component rotation.

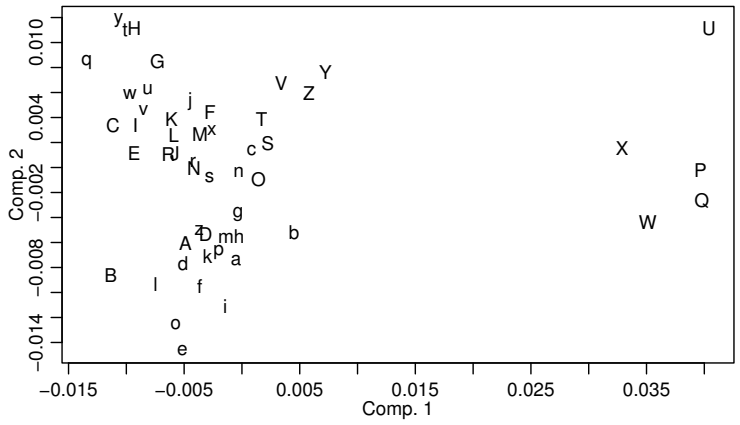

Figure 5.3.b: (i) Scatterplot of the first two principal components for the example "NIR-spectra".

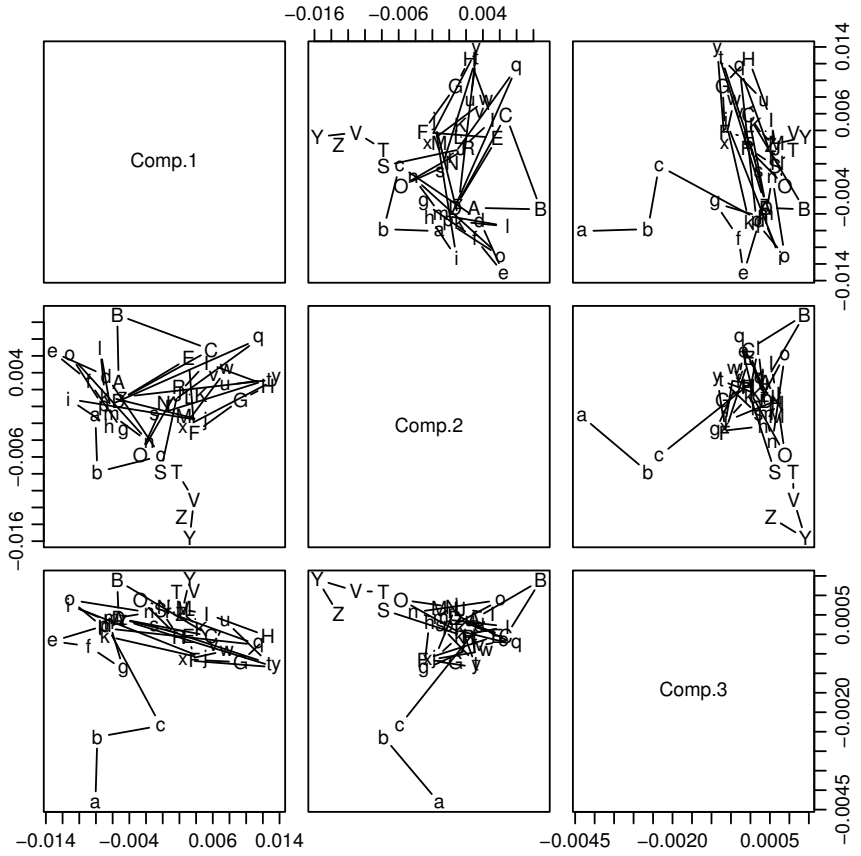

Figure 5.3.b: (ii) Scatterplot matrix of the first three principal components for the example "NIR-spectra" without the 5 outliers.

c PCA is suitable for many multivariate data sets. If we are analyzing spectra we have the special case that the variables (the intensities of different wavelengths) have a special ordering. Hence, we can plot each observation as a "function". We can also illustrate the principal component directions (the loadings) $\underline{b}_k$ as spectra!

d **Scaling Issues** If the variables are measured in different units, they should be standardized to (empirical) variance 1 (otherwise comparing variances doesn't make sense). This leads to a PCA (= eigenanalysis) of the correlation- instead of the covariance matrix.

For spectra this is *not* useful because wavelengths with very variable intensities contain the most important information. If we would standardize the variables in that setup, we would down-weight these variables compared to the unstandardized data set.

e **Choosing the number p of components**: $(p < m)$

- 2 (maybe 3) for illustrational purposes.
- Plot the explained variance (eigenvalues) in decreasing order and look for a breakpoint (**"scree plot"**: plot $\widehat{\lambda}_k$ vs. k), see Figure 5.3.e.
- "Explain 95% of the variance": The sum of the eigenvalues $\sum_{j=1}^{p} \widehat{\lambda}_j$ should be 95% of the total sum $\sum_{j=1}^{m} \widehat{\lambda}_j$.

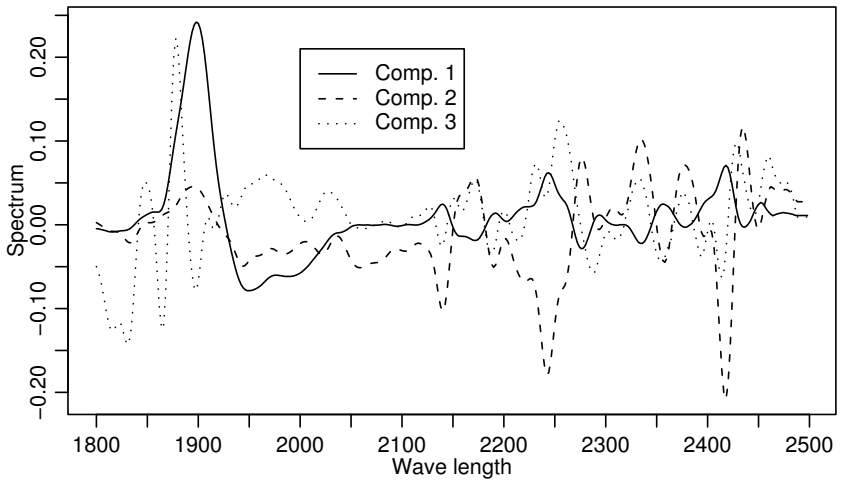

Figure 5.3.c: Spectra of "loadings" of the first three principal components for the example "NIR-spectra".

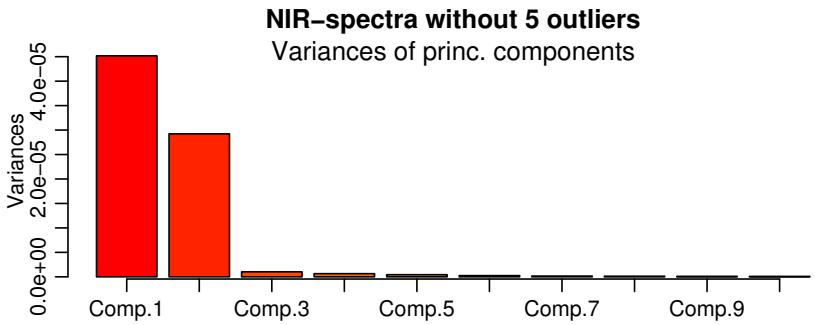

Figure 5.3.e: Variances of the principal components (scree plot) for the example "NIR-spectra".

But: "Variance" $\sum_{j=1}^{m} \lambda_j = \sum_{j=1}^{m} \mathrm{Var}(X^{(j)})$ is the sum of all variances. There could be (many) noise variables among them!

Restriction to the first p principal components: In the transformation formula (5.3.a) we simply ignore the last $m - p$ terms:

$$\underline{X}_i - \widehat{\underline{\mu}} = \widehat{\underline{X}}_i + \widehat{\underline{E}}_i \ , \quad \widehat{\underline{X}}_i = \sum_{k=1}^{p} Z_i^{(k)} \underline{b}^{(k)} \ , \quad \widehat{\underline{E}}_i = \sum_{k=p+1}^{m} Z_i^{(k)} \underline{b}^{(k)}.$$

This can be interpreted in the following two ways.

- In Linear Algebra terminology:
  The "data matrix" of the $\widehat{\underline{X}}_i$ is the best approximation of the data matrix of the $\underline{X}_i - \widehat{\underline{\mu}}$ if we restrict ourselves to matrices with rank p (in the sense of the so-called Frobenius norm of matrices: $\|\boldsymbol{E}\|^2 = \sum_{ij} E_{ij}^2$).
- In statistical terminology:
  We were looking for p variables $Z^{(k)} = \sum_{j=1}^{m} B_{kj} X^{(j)}$, $k = 1, \ldots p$, such that the differences $\underline{E}_i = \underline{X}_i - \widehat{\underline{X}}_i$ of $\widehat{\underline{X}}_i = \sum_{k=1}^{p} Z_i^{(k)} \underline{b}^{(k)}$ show minimial variance (in the sum): $\sum_{j=1}^{m} \widehat{\mathrm{Var}}(E^{(j)}) = \sum_{k=p+1}^{m} \lambda_k$ is minimal (there will be no better choice than the variables $Z^{(k)}$).

5.4 Linear Mixing Models, Factor Analysis

a **Model for Spectra** Let $\underline{c}_k$ be the spectrum of the chemical component k and consider a mixture of the components with coefficients $\underline{s} = [s^{(k)}]$. For the ith mixture we have the coefficients $\underline{s}_i$. According to Lambert-Beer the spectrum of the ith mixture is

$$\underline{X}_i = \sum_k \underline{c}^{(k)} \underline{s}_i^{(k)} + \underline{E}_i = \boldsymbol{C}\underline{s}_i + \underline{E}_i$$

where $\underline{E}_i$ are measurement errors. $\boldsymbol{C}$ is the matrix of spectra $\underline{c}_k$ (in the different columns).

This looks very similar to 5.3.e. The differences are

- $\boldsymbol{C}$ not orthogonal
- $\underline{X}_i$ instead of $\underline{X}_i - \underline{\widehat{\mu}}$, not centered
- $\underline{E}_i$ random vector (measurement error)
- $s_i^{(k)} \geq 0$ or $C_{jk} \geq 0$, $X_i^{(j)} \geq 0$ if we use the original spectra.

b This model can be used for many applications where there are m measurements that are linear superimpositions of $p < m$ components.

Examples are:

- Chemical elements in rocks that consist of several bed-rocks.
- Trace elements in spring water that ran through different soil layers.

c If the source profiles (spectra) $\underline{c}_k$ are known, the "contributions" $s_i^{(k)}$ can be estimated for each observation i separately using linear regression.

However, it's more interesting if both the source profiles and their contributions have to be estimated from data. This can be achieved using a combination of statistical methods, professional expertise and application specific properties.

5.5 Regression with Many Predictors

a In the introductory example about NIR-spectra we discussed the question whether we can "predict" the amount of an active ingredient based on a spectrum.

Hence, we have a response variable Y and several predictors $[x^{(1)}, ..., x^{(m)}]$. If we set up a linear regression model we face the problem that there are many more predictors than observations. Hence, it's not possible to fit a "full model" (it would lead to a perfect fit).

A possible remedy is to use **"stepwise"** regression: We start with just one predictor and add the most significant predictor in the next step (until some stopping criterion is met).

Example: Granulate Samples.
Y = yield. $n = 44$ (without "outliers"). Table 5.5.a shows a computer output. For comparison: Simple correlation between L2450 and yield: $r = -0.57$, $R^2 = 0.32$.

b Better known are the following methods to handle the problem of having too many predictors

1. **Principal Component-Regression**,
2. Ridge Regression
3. New methods like Lasso, Elastic Net, ...

```
                  Value  Std. Error  t value  Pr(>|t|)  Signif
(Intercept)    75.30372     0.07175  1049.52     0.000  ***
L2450        -395.43390    76.70623    -5.16     0.000  ***
L2010        -465.28939   142.44458    -3.27     0.002  **
L1990         585.20468   128.49676     4.55     0.000  ***
L2360         875.33702   160.04160     5.47     0.000  ***
L2400         532.91971   117.74430     4.53     0.000  ***
L2480        -301.44225    77.70208    -3.88     0.000  ***
L2130        -501.39852    88.17596    -5.69     0.000  ***

Residual standard error: 0.2268 on 36 degrees of freedom
Multiple R-Squared: 0.7212
```

Table 5.5.a: Computer output for a regression model after variable selection with stepwise forward.

c **Principal Component-Regression** PCA of the predictors leads to new variables $[Z^{(1)}, \ldots, Z^{(p)}]$. The principal components are usually selected without examining the relationship with the response Y.
Variant of Brown Brown (1993): Select them according to simple correlation with Y!

d **Ridge Regression** An easy way to ensure that the matrix $\boldsymbol{X}^T\boldsymbol{X}$ (that needs to be invertible for least squares) is non-singular is to add a diagonal matrix $\lambda\boldsymbol{I}$, leading to

$$\underline{\widehat{\beta}}_\lambda = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\underline{Y}.$$

Bibliography

Bates, D. M. and Watts, D. G. (1988). *Nonlinear regression analysis and its applications*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Bennett, J. H. (ed.) (1971). *Collected Papers of R. A. Fischer; 1912-24*, Vol. I, The University of Adelaide.

Boen, J. R. and Zahn, D. A. (1982). *The Human Side of Statistical Consulting*, Wadsworth Inc. Belmont.

Bortz, J. (2005). *Statistik für Sozialwissenschaftler*, 6 edn, Springer, Berlin.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978). *Statistics for Experimenters*, Wiley, N. Y.

Brown, P. J. (1993). *Measurement, Regression, and Calibration*, Clarendon Press, Oxford, UK.

Carroll, R. and Ruppert, D. (1988). *Transformation and Weighting in Regression*, Wiley, New York.

Chatfield, C. (1996). *The Analysis of Time Series; An Introduction*, Texts in Statistical Science, 5 edn, Chapman and Hall, London, NY.

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*, Wiley Series in Probability & Mathematical Statistics, Wiley, New York.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*, 2 edn, Wiley, N. Y. 1st ed. 1971.

Federer, W. T. (1972, 1991). *Statistics and Society: Data Collection and Interpretation*, Statistics: Textbooks and Monographs, Vol.117, 2 edn, Marcel Dekker, N.Y.

Harman, H. H. (1960, 1976). *Modern Factor Analysis*, 3 edn, University of Chicago Press, Chicago.

Hartung, J., Elpelt, B. and Klösener, K. (2002). *Statistik. Lehr- und Handbuch der angewandten Statistik*, 13 edn, Oldenbourg, München.

Hoaglin, D. C., Mosteller, F. and Tukey, J. W. (1991). *Fundamentals of Exploratory Analysis of Variance*, Wiley, N. Y.

Hogg, R. V. and Ledolter, J. (1992). *Applied Statistics for Engineers and Physical Scientists*, 2 edn, Maxwell Macmillan International Editions.

Huet, S., Bouvier, A., Gruet, M.-A. and Jolivet, E. (1996). *Statistical Tools for Nonlinear Regression: A Practical Guide with S-Plus Examples*, Springer-Verlag, New York.

Lawley, D. N. and Maxwell, A. E. (1963, 1967). *Factor Analysis as a Statistical Method*, Butterworths Mathematical Texts, Butterworths, London.

Linder, A. and Berchtold, W. (1982). *Statistische Methoden II: Varianzanalyse und Regressionsrechnung*, Birkhäuser, Basel.

Mead, R. (1988). *The design of experiments*, Cambridge University Press, Cambridge.

Myers, R. H. and Montgomery, D. C. (1995). *Response Surface Methodology; Process and Product Optimization Using Designed Experiments*, Wiley Series in Probability and Statistics, Wiley, NY.

Petersen, R. G. (1985). *Design and Analysis of Experiments*, Statistics Textbooks and Monographs, Marcel Dekker, N.Y.

Rapold-Nydegger, I. (1994). *Untersuchungen zum Diffusionsverhalten von Anionen in carboxylierten Cellulosemembranen*, PhD thesis, ETH Zurich.

Ratkowsky, D. A. (1989). *Handbook of Nonlinear Regression Models*, Marcel Dekker, New York.

Renner, R. M. (1993). The resolution of a compositional data set into mixtures of fixed source compositions, *Applied Statistics — Journal of the Royal Statistical Society C* **42**: 615–631.

Sachs, L. (2004). *Angewandte Statistik*, 11 edn, Springer, Berlin.

Scheffe, H. (1959). *The Analysis of Variance*, Wiley, N. Y.

Seber, G. and Wild, C. (1989). *Nonlinear regression*, Wiley, New York.

Stahel, W. A. (2000). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 3 edn, Vieweg, Wiesbaden.

Swinbourne, E. S. (1971). *Analysis of Kinetic Data*, Nelson, London.