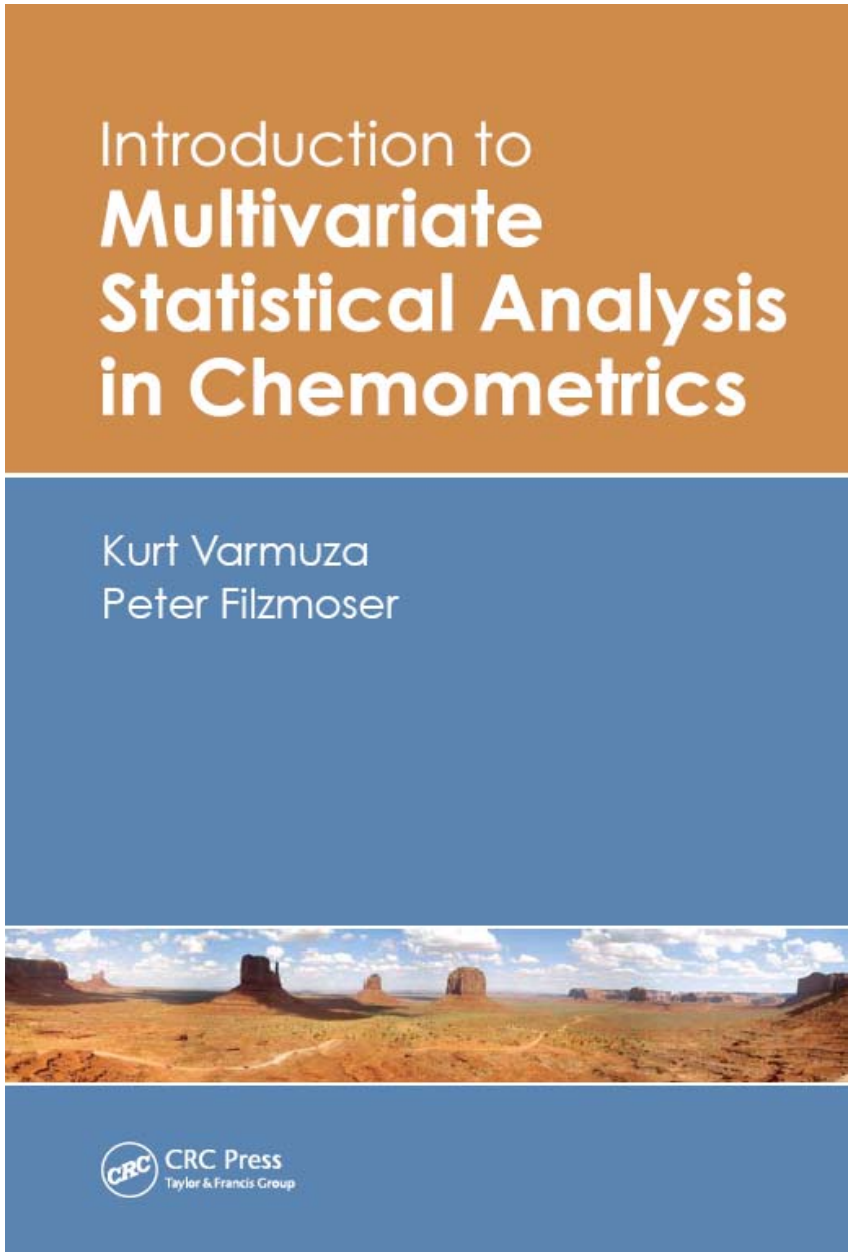


Authors

Kurt Varmuza
Peter Filzmoser

Vienna University of Technology, Austria

kurt.varmuza@tuwien.ac.at
peter.filzmoser@tuwien.ac.at

Info: www.lcm.tuwien.ac.at

ISBN: 9781420059472

CRC Press (Taylor & Francis),
Boca Raton, FL, USA, www.crcpress.com

2009

321 + xiii pages

Ca US$ 180

(220129)

- Includes important multivariate statistical methods, such as principal component analysis and support vector machines, for analyzing scientific data
- Explains the methods using formulae, graphical illustrations, and schemes
- Demonstrates **R software tools** with fully worked-out, real-world examples
- Emphasizes the use of **robust statistical methods**
- Offers practical advice on applying the methods

Using formal descriptions, graphical illustrations, practical examples, and R software tools,

Introduction to Multivariate Statistical Analysis in Chemometrics

presents simple yet thorough explanations of the most important multivariate statistical methods for analyzing chemical data.

It includes discussions of various statistical methods, such as principal component analysis, regression analysis, classification methods, and clustering.

Written by a chemometrician and a statistician, the book reflects both the practical approach of chemometrics and the more formally oriented one of statistics. To enable a better understanding of the statistical methods, the authors apply them to real data examples from chemistry. They also examine results of the different methods, comparing traditional approaches with their robust counterparts. In addition, the authors use the freely available R package to implement methods, encouraging readers to go through the examples and adapt the procedures to their own problems.

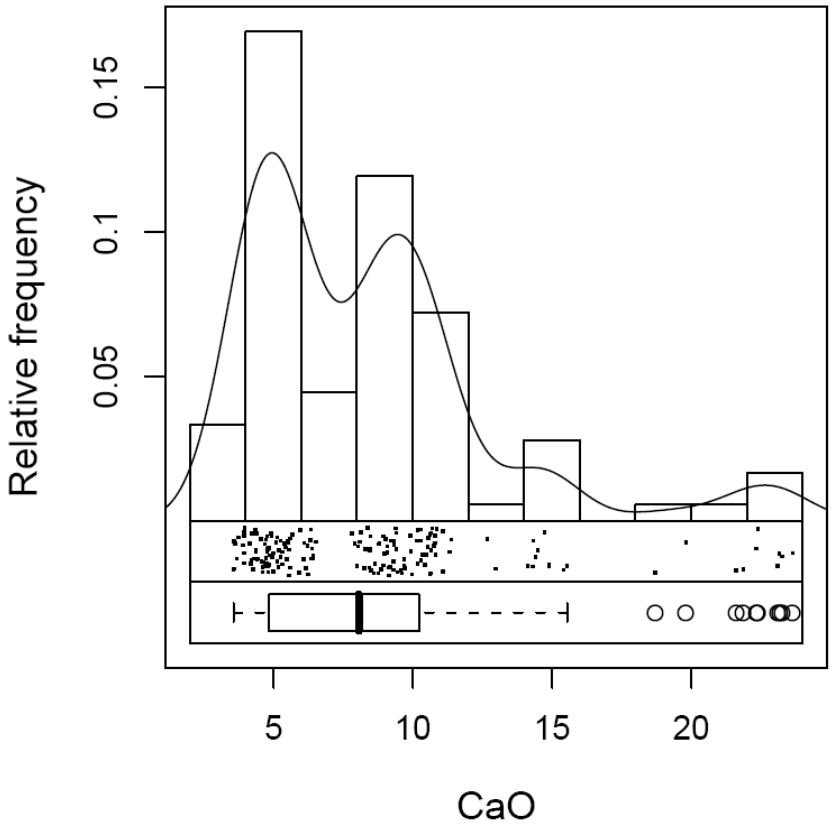
Focusing on the practicality of the methods and the validity of the results, this book offers concise mathematical descriptions of many multivariate methods and employs graphical schemes to visualize key concepts. It effectively imparts a basic understanding of how to apply statistical methods to multivariate scientific data.

Contents

Examples from the book

1. Introduction

1.6. Univariate Statistics - A Reminder



The BOXPLOT is an informative graphics to display a data distribution, based on median and quartiles. The boxplot can be defined as follows [Frank and Todeschini 1994]. The height of the box is given by the first and third quartile, and the mid line shows the median; the width of the box has usually no meaning. One whisker extends from the first quartile to the smallest data value in the interval Q_1 to $Q_1 - 1.5 \cdot IQR$ and is called the lower whisker. The other whisker extends from the third quartile to the largest data value in the interval Q_3 to $Q_3 + 1.5 \cdot IQR$ and is called the upper whisker. Outliers - not within the range $[Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR]$ - are plotted as individual points.

Figure 1.7. The EDAPLOT (Exploratory Data Analysis plot) of data combines one-dimensional scatter plot, histogram, probability density trace, and boxplot. Data used are CaO concentrations (%) of 180 archaeological glass vessels [Janssen *et al.* 1998].

Examples from the book

2. Multivariate Data
2.6. Linear Latent Variables

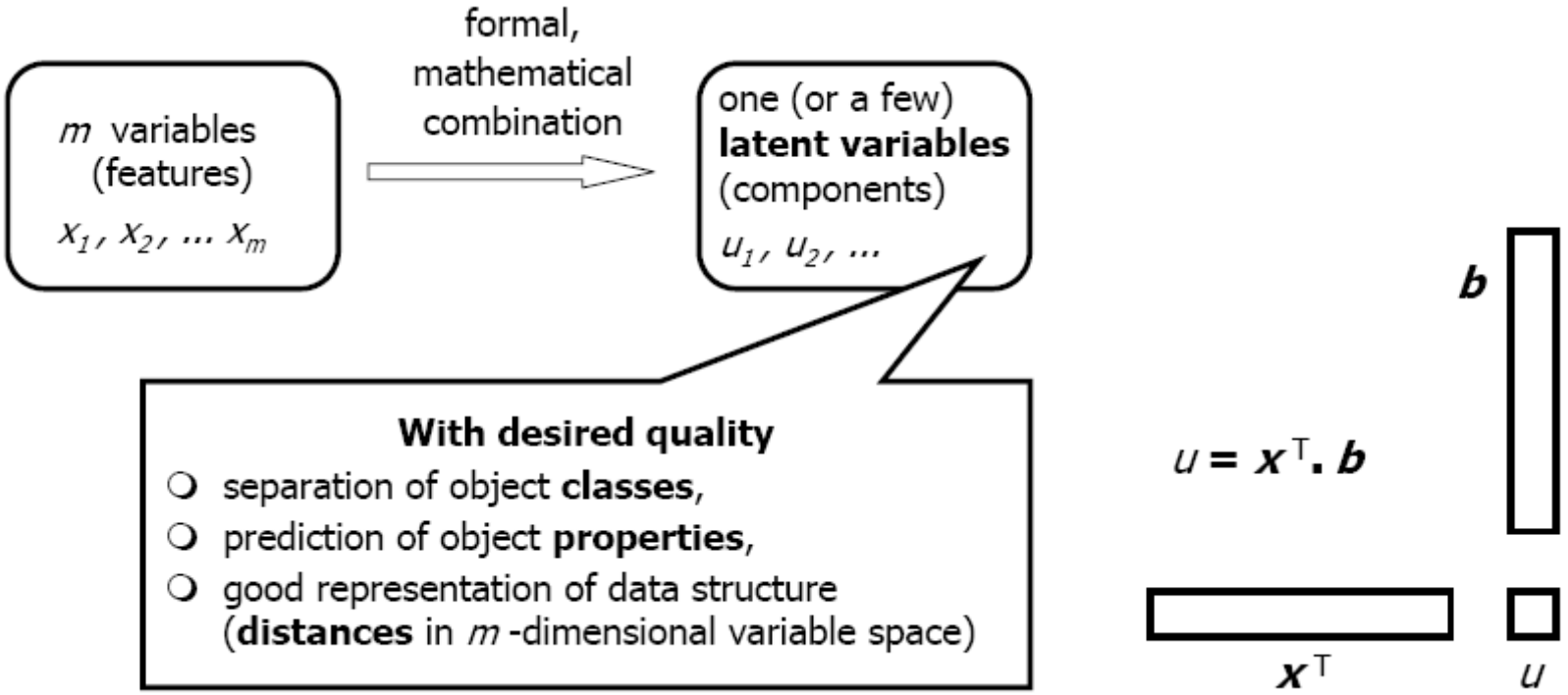


Figure 2.14. General concept of a latent variable (left), and calculation of the score u of a linear latent variable from the transposed variable vector $\boldsymbol{x}^T$ and the loading vector $\boldsymbol{b}$ as a scalar product (right).

Examples from the book

3. Principal Component Analysis

3.1. Concepts

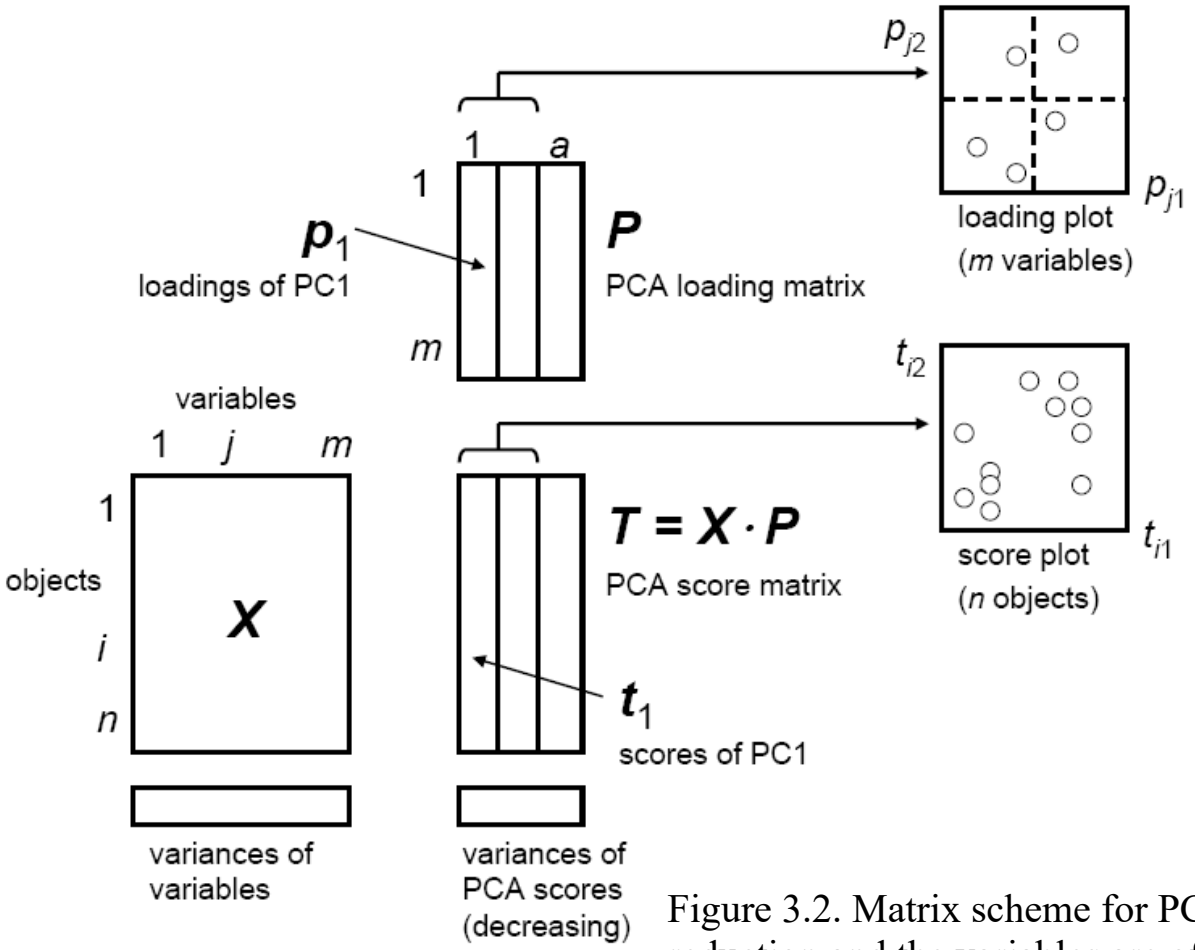


Figure 3.2. Matrix scheme for PCA. Since the aim of PCA is dimension reduction and the variables are often highly correlated $a \leq \min(n,m)$ principal components are used.

Principal component analysis (PCA) can be considered as "the mother of all methods in multivariate data analysis". The aim of PCA is dimension reduction and PCA is the most frequently applied method for computing linear latent variables (components). PCA can be seen as a method to compute a new coordinate system formed by the latent variables, which is orthogonal, and where only the most informative dimensions are used. Latent variables from PCA optimally represent the distances between the objects in the high-dimensional variable space - remember, the distance of objects is considered as an inverse similarity of the objects.

Examples from the book

4. Calibration
4.9. Examples

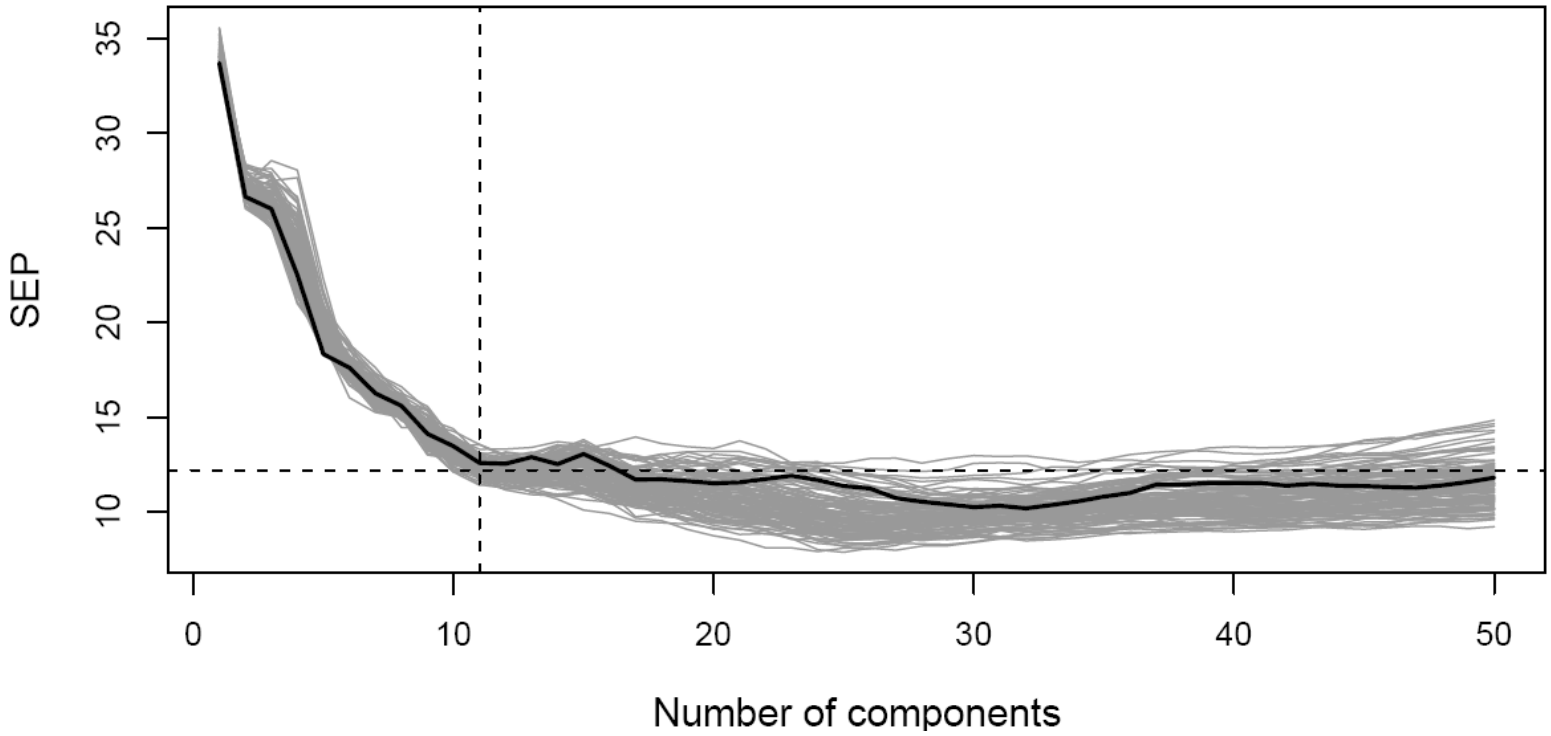


Figure 4.34. Results of PLS for the PAC data set. The black line results from a single 10-fold cross validation, the gray lines from repeating the 10-fold cross validation 100 times. The choice of the optimal number of principal components is based on repeated double cross validation.

Examples from the book

5. Classification
5.3.3. k-NN Classification

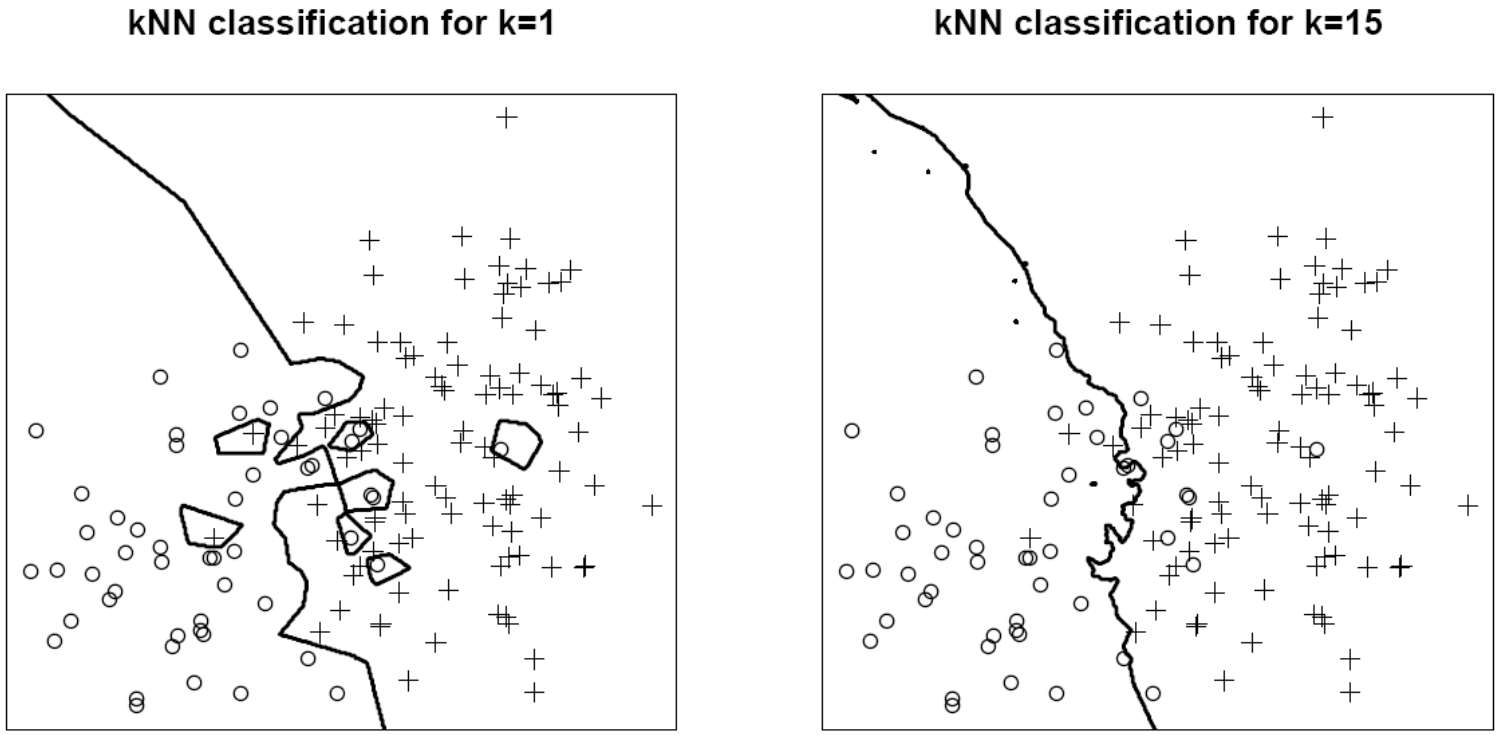


Figure 5.13. kNN classification for two groups of 2-dimensional data. The training data are shown with the symbol corresponding to the group membership. Any new data point would be classified according to the presented decision boundaries, where $k = 1$ in the left plot, and $k = 15$ in the right plot has been used.

Examples from the book

6. Cluster Analysis

6.1. Concepts

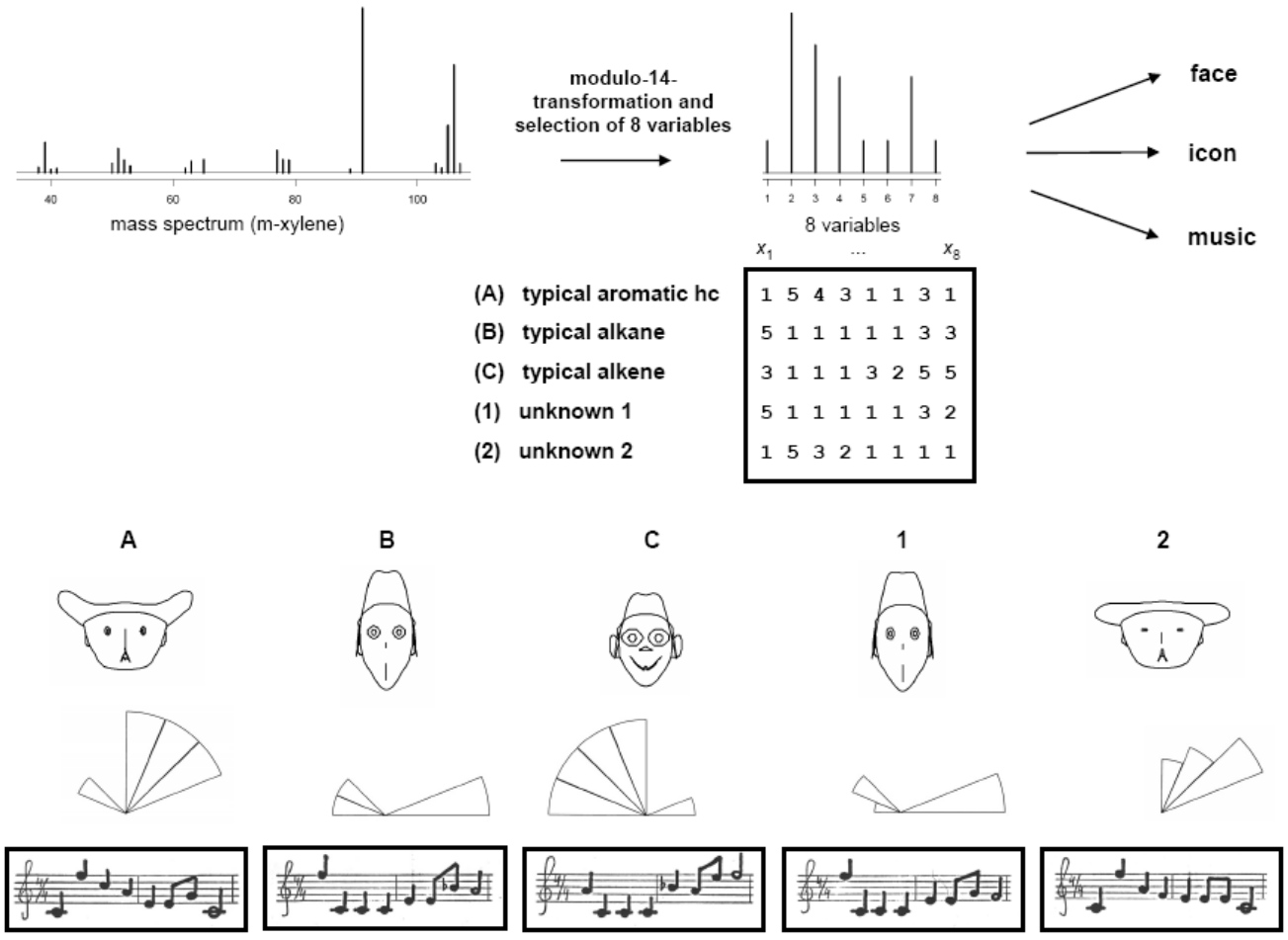


Figure 6.2. Representation of multivariate data by icons, faces and music for human cluster analysis and classification in a demo example with mass spectra. Mass spectra have first been transformed by modulo-14 summation (see Section 7.4) and from the resulting 14 variables 8 variables with maximum variance have been selected and scaled to integer values between 1 and 5. **A**, typical pattern for aromatic hydrocarbons; **B**, typical pattern for alkanes; **C**, typical pattern for alkenes; **1** and **2**, "unknowns" (2-methyl-heptane and meta-xylene). The 5×8 data matrix has been used to draw faces (by function "faces" in the ℝ-library TeachingDemos), segment icons (by ℝ-finction "stars"), and to create small melodies [Varmuza 1986]. Both unknowns can be easily assigned to the correct class by all three representations.