Applied Multivariate Statistics

Course 701-0102-00, "with Complements" 401-0102-99

and Advanced Studies Course in Applied Statistics

Spring Semester 2011, ETH Zurich

Werner Stahel

Seminar for Statistics, ETH Zurich

Translated to English by Armanda Strong

February-May, 2011

written consent from the author.

1 Introduction

1.1 Questions of Multivariate Statistics

a For characterizing people, objects, or other observed entities, one or many features are usually recorded. If blood pressure, age, sex, weight, treatment type, and other data are collected for patients, then the interest usually lies in representing the variable of interest – blood pressure – as a function of explanatory variables to capture causal relationships like the effectiveness of the treatment. This leads to the underlying question of statistical regression methodology. However, often **multiple features are of equal interest**: For insect larva, the length of multiple limbs are measured; for patients, upper and lower blood pressures are obtained, as well as the concentration of multiple substances in the blood or other measurable values are important; for a chemical reaction, the concentrations of multiple agents are involved, etc.

b **Multivariate Statistics** deals with situations where multiple variables should be considered jointly and with equal importance.

c For the **graphical representation** *of an individual* variable, variants of histograms and box plots are of primary use. For reflecting the joint distribution of *two* variables, the scatter plot (Fig. 1.2.b(i)) is the "basic figure". The **joint distribution of *multiple* variables** is harder to represent, and so the **graphical methods** in multivariate statistics are based on a variety of ideas.

d The basis of statistics with a single variable or a singe variable of interest in regression is formed by **probability distributions**. For understanding the relationships between multiple variables, it is more important to develop **probability models** that can describe the relationships. On this basis, we can answer the **basic questions of inferential statistics** about the relationship between observed data and parameter values of a model. Finally, it will be important to evaluate methods for checking model assumptions.

e When considering a single random variable, the issues of **estimating and testing a mean** and the **comparison of two or more groups** are basic problems that stand behind inferential statistics. They are extended to more variables in multivariate statistics, and similarly, the normal **variance analysis** and **regression** with one variable of interest is generalized to the joint consideration of **multiple variables of interest**.

f In multivariate statistics, we repeatedly start with questions that are obvious when studying a single random variable and generalize the problem and the corresponding methods to the case of multiple variables. To the simpler case we therefore give the name **univariate statistics**.

1.2 Examples

a ▷ **Iris Species.** In the year 1935, the biologist E. Anderson measured the length and width of the sepal leaves and petal leaves of various species of iris flowers; both leaf types are part of the flowers. Sir R. A. Fisher (1936) used the data from Anderson as a basic example for a multivariate procedure, "discriminant analysis"; because of this, this data set became probably the most famous in all of statistics. Of the 150 plants whose measurements Fisher used, there were 50 of each of the species Iris setosa, Iris virginica, and Iris versicolor.

The question of the study is whether, purely on the basis of the measured flower leaves, the **plants can be allocated to the individual species**. It is suspected that Iris versicolor is really a hybrid of the two other species. (This information about the data set is taken from Andrews and Herzberg (1985).)

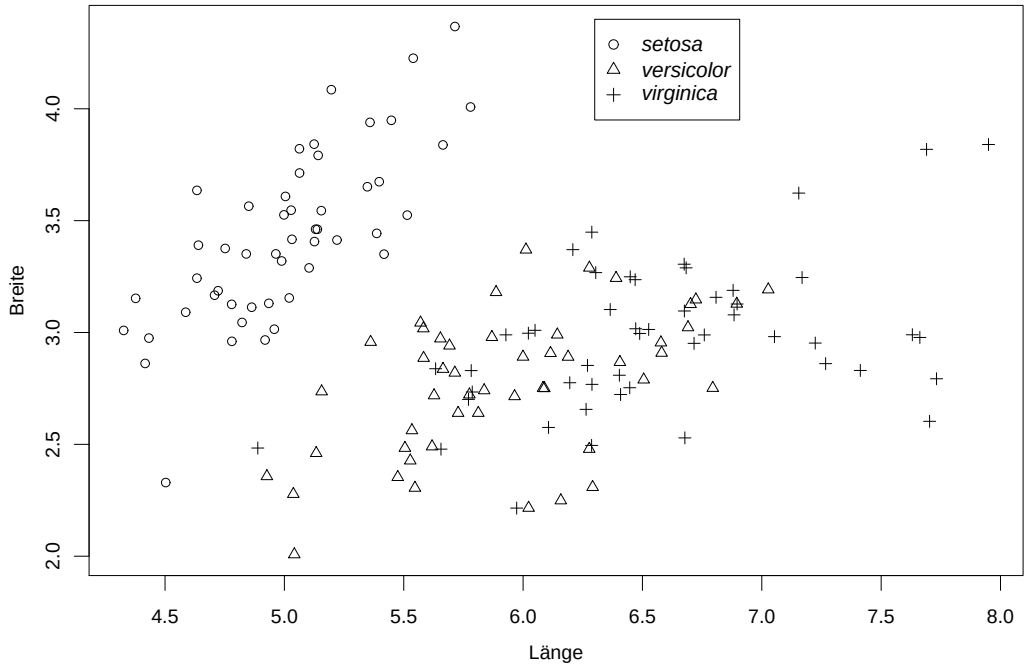

Figure 1.2.b (i): Scatter plot of the length and width of the sepal leaves in the iris flowers example

b ▷ We initially consider only two variables, the length and width of the sepal leaves. It is obvious to draw the data into a scatter plot (Fig. 1.2.b (i)). The points, which correspond to the individual plants, are marked by symbols that reflect the species.

We see that – as expected – the two variables are related to each other; most longer leaves are also wider than shorter ones. We speak of correlation of the two variables. It is also obvious that the joint consideration of these correlated variables is more helpful than the separate representation (Fig. 1.2.b (ii)): In the joint representation, the species "setosa" can be clearly separated from the other two, which is not possible for either the lengths nor the widths alone. ◁

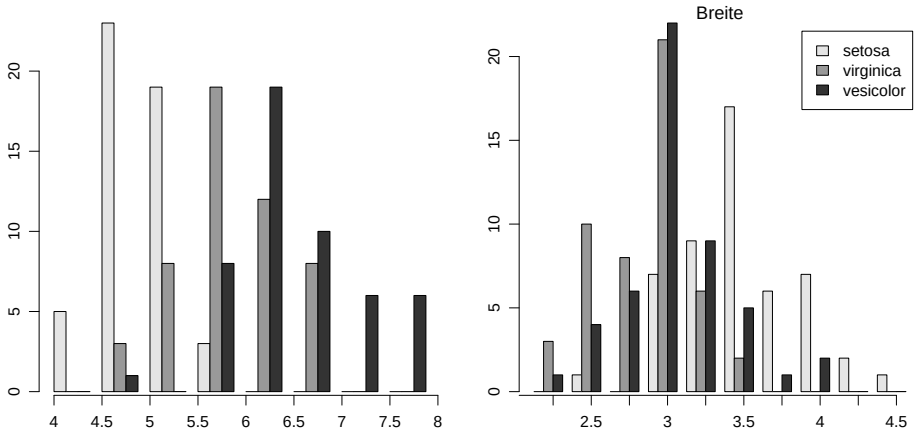

Figure 1.2.b (ii): Histograms of lengths and widths of the sepal leaves, divided into the three species Iris setosa, virginica, and versicolor

c ▷ **Vein Constriction. Medical diagnostic tests** serve to divide patients into sick and healthy with respect to a specific illness. Often this is done on the basis of a measurement of an individual variable or even a yes-no answer. However, a much more accurate diagnosis is possible if multiple symptom features are used simultaneously.

A simple example is the diagnosis of a vein constriction on the basis of the heartbeat volume (Vol) and the pulse (Rate). Fig. 1.2.c shows the data after a log transformation. (Sources: Finney, 1947, Biometrika 34 and also Fahrmeir, Hamerle and Tutz, 1996). ◁

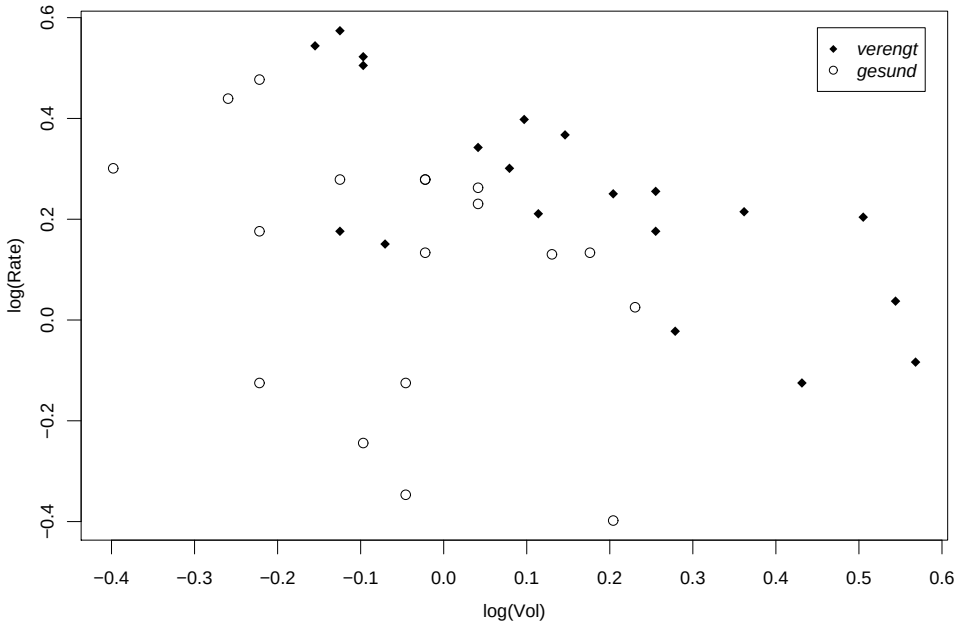

Figure 1.2.c: Data in the vein constriction example

d ▷ **Fossils.** From fossils that are found in different layers of the seabed, we want to infer environmental conditions (temperature, nutrient content) for the corresponding time period.

?) therefore collected measurements of various morphological features of cocoliths of the species *Gephyrocapsa* at 110 points in the oceans in the uppermost (holocene) layer. Fig. 1.2.d shows these points and a schematic diagram of a cocolith. The values lie in the region of 1-5 μm. At the sample points, the current environmental conditions are also collected.

Additionally, cocoliths from deeper strata were measured. The basic idea is that the morphological features depend on the environmental conditions. The relationship can be modeled on the basis of relating the contemporary samples to today's climate. If the relationships have remained the same, we can use the **morphological features of the deeper layers to infer the environmental conditions at that time**.

The research group found that better prognoses resulted if the individuals were first divided into subspecies with help of the mentioned measurements and then the proportions of the subspecies in the sample were used. The question thus arises how we should **define such subspecies** and how we should assign the individuals to these subspecies. Such a definition is indicated in subfigure C. ◁

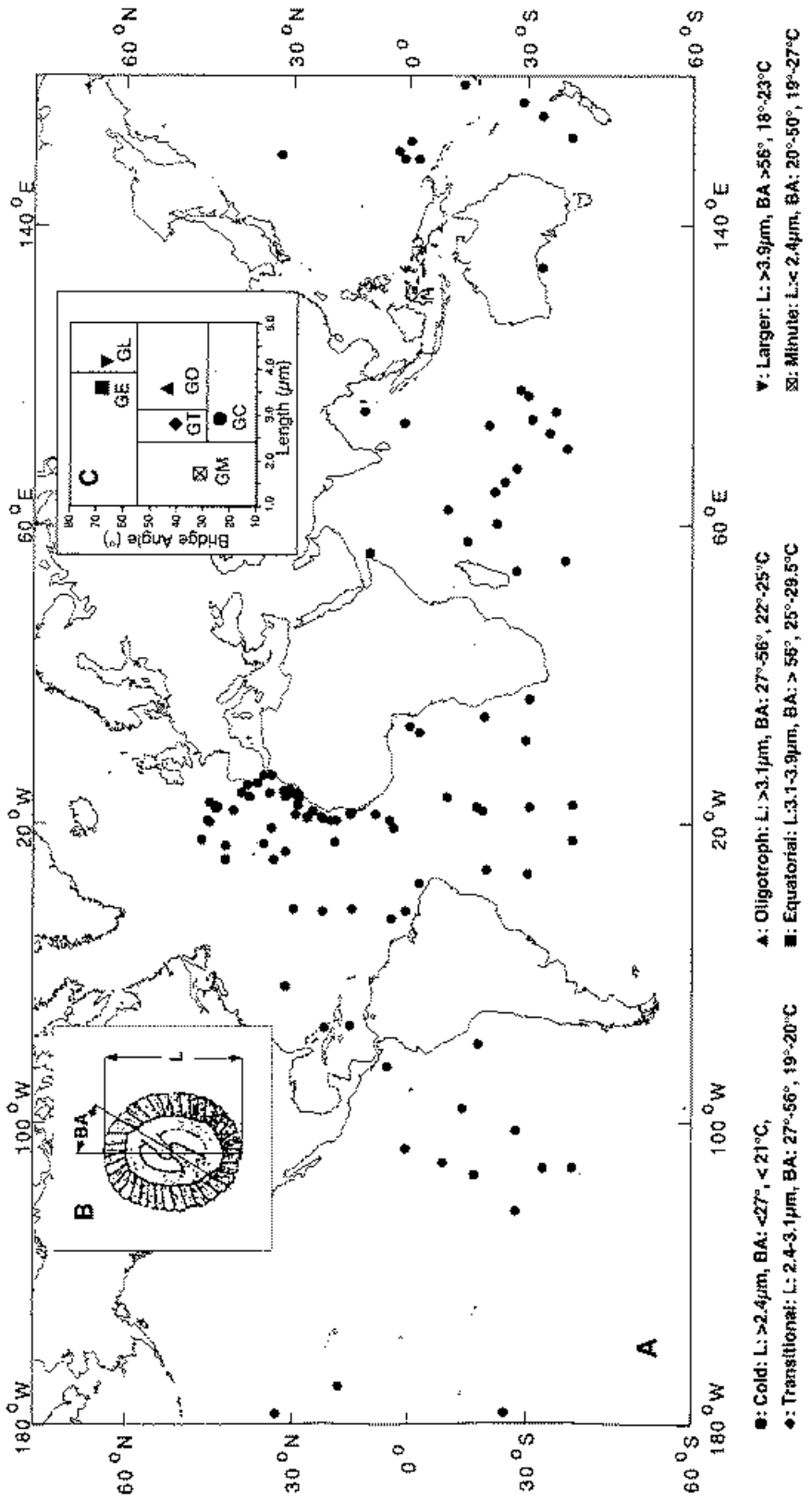

Figure 1.2.d: Sampling locations in the fossil example (A). (B) For the basic shape of the cocoliths, two important measurements of the length and the "bridge angle" are give. (C) shows a classification into subspecies on the basis of the two measurement values.

e ▷ **Ecosystem.** In studies about ecosystems, the relationship between various types of variables is studied. **?**) posed the question of how grazing influences vegetation. On an alp in the canton of Ticino, for 82 sample plots they first determined the intensity of grazing via observation of the location type and of excrements found, then they determined the values of chemical properties – pH, phosphate, nitrate, and carbon concentrations – and the frequency of 64 plant species. The location of the sample determines physical environmental variables like slope, exposure, and height above sea level.

Fig. 1.2.e shows how the abundance of the 6 most frequent species depends on the grazing intensity. How can we characterize the vegetation as a whole without having to study all 64 species individually? How strongly do the variable groups depend on each other? ◁

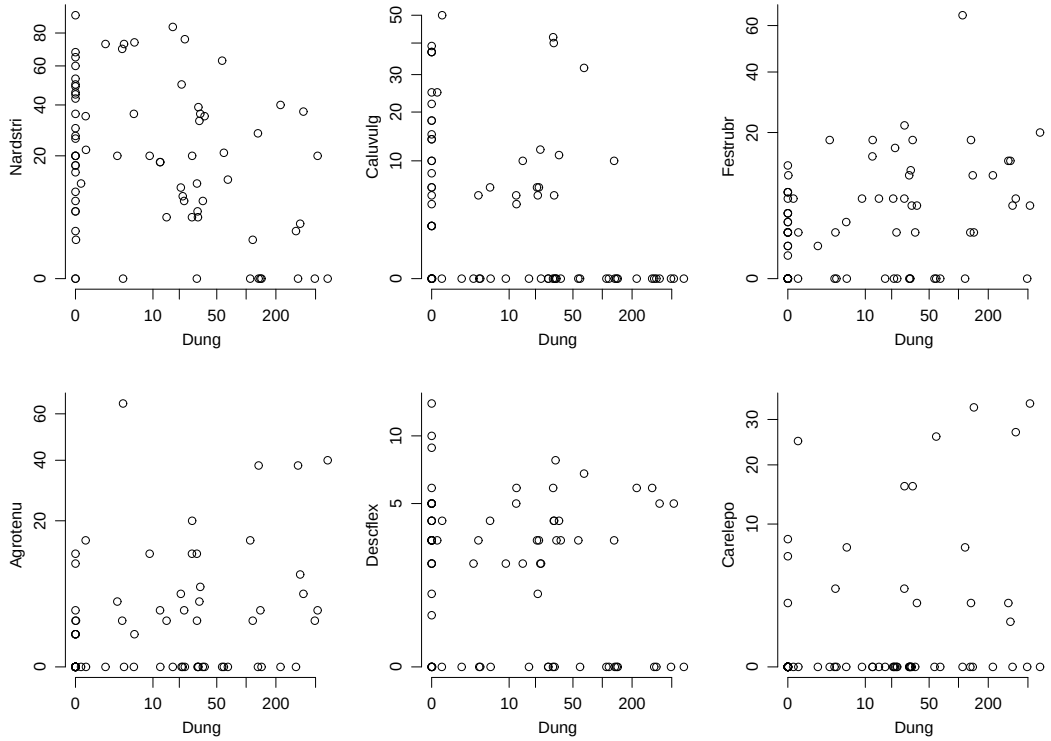

Figure 1.2.e: Dependency of the 6 most frequency plant species on the grazing intensity in the ecosystem example

f ▷ **Voting.** With statistical analysis of voting results, you can make it into the newspapers. For 14 federal issues of the year 1995-96, the proportion of yes votes for the cantons were collected. Are interpretable patterns apparent therein?

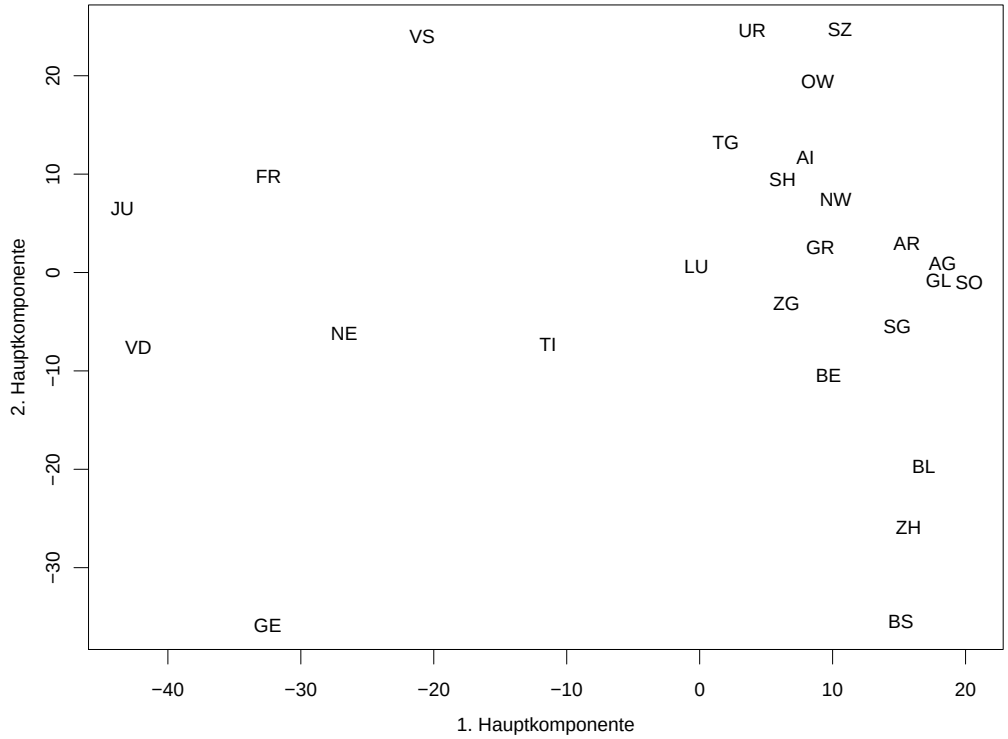

Figure 1.2.f: First two principal components in the voting example

Fig. 1.2.f shows a **graphical representation of the result** from a multivariate analysis, known as Principal Component Analysis, of the proportions. In the left half we see cantons from western Switzerland, on the right the cantons in eastern Switzerland, and in the middle the southern Ticino. On the top are the rural cantons and on the bottom, the urban. This "geography" was not used to generate the representation. Obviously, similarities in voting behavior are enough to find these contrasts. ◁

g ▷ **NIR Spectra.** Spectra play an important role in chemistry. They allow the **composition of mixtures** to be determined without chemical analysis, that is, without carrying out a specific reaction with a small sample for each substance in question to find its concentration. There are many types of spectra. Optical spectra measure the absorption of light that is beamed through a sample (or reflected by a solid sample) in relationship to the wavelengths of the light.

In the ideal case, for each substance in the mixture there is a "peak" in the spectrum – for a certain wavelength this substance absorbs a lot of light, but for other wavelengths negligibly little – and these characteristics wavelengths are distinguishable for different substances so that peaks don't overlap. The analysis of such spectra is relatively easy: The size (area) of the peaks is directly proportional to the proportion of the corresponding substance in the mixture.

For wavelengths in the region of near infrared (NIR), this is unfortunately not the case. The individual substances absorb NIR spectra over larger areas and the "peaks" are blurred and strongly overlapping. It is better to replace the idea of peaks with the a specific spectrum of a general form for each substance.

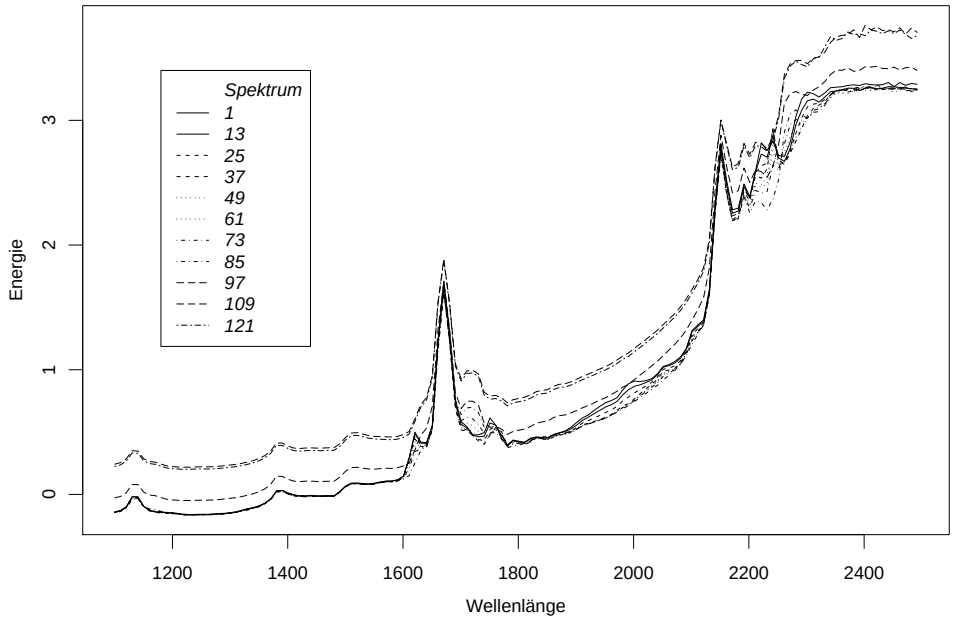

Figure 1.2.g (i): 11 NIR spectra that document the course of a chemical reaction

We can deal with the absorption of a mixture for a certain wavelength as a variable. For each wavelength for which the absorption is measured we get a variable. Fig. 1.2.g (i) shows 11 of 121 NIR spectra recorded in the course of a chemical reaction.

Multivariate statistics is involved since, from the joint information from these(arbitrarily) many variables, we extract, if possible, the concentrations of the substances in each mixture. Here, a law of nature is beneficial: The spectrum of a mixture in the weighted sum of the spectra of the individual substances; the weights are the proportions of the substances in the mixture. Not surprisingly, this rule is not exact, but only an approximation and is valid only up to measurement error. It is often called the Lambert-Beer law (even though the actual Lambert-Beer law only describes the relationship of the intensity with the characteristic values of the measurement equipment).

With this basic law and corresponding multivariate methods, reactions can be researched or processes supervised in chemistry. With it's help, in the example, 4 phases of the process can be identified that occur one after the other; Fig. 1.2.g (ii) shows the sequence of these phases over time.

We will describe the law with a probability model that is known as **the linear mixing model** . It also has applications in the analysis of environmental pollutants and other areas. ◁

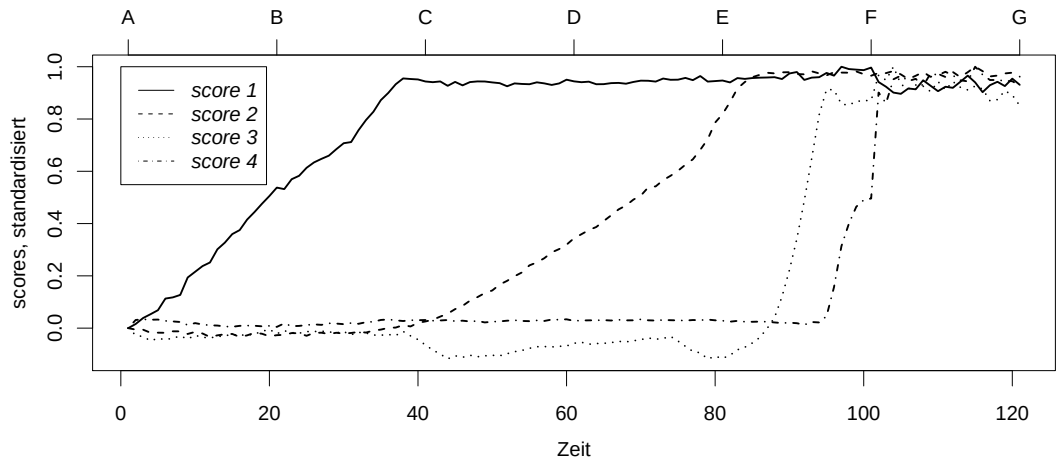

Figure 1.2.g (ii): Sequence in time of four phases of the process in the NIR spectra example

h **Customer Segmentation.** An important application field for multivariate data is the ever-expanding pool of data about customers of large firms. For marketing purposes, customers with specials interests or behavior should be found so that they can be reached with targeted advertising or otherwise specially treated.

In certain applications we use data from the past to derive rules for prediction. For example, the data from a bank's debtors is used to predict which of them have a higher probability of becoming bankrupt. To derive the rule, we use data from the past and compare the groups of those who become bankrupt with those that did not have problems. We might also want to predict the revenues from individual customers. This would be a question of regression.

In other applications there is no definite variable of interest or classification that we want to predict, instead we would like to find "reasonable groups" in the data, which might help to focus advertizing campains.

1.3 Questions

a The presented examples illustrate the following questions that can be answered with the methods of multivariate statistics .

- To begin with, using **graphical representations** we would like to make a picture of the data, whether for **screening** of the data, for **explorative analysis**, for searching for unexpected features or, later, for illustration of relationships that have resulted from analysis with other methods.

b
- If we consider the joint distribution of two variables, then we can ask: **Are the variables independent?** If not: **Which type of dependency is present? How strong is it?** The most well-known way of characterizing a relationship is the (Pearson) correlation coefficient, which measures the strength of a linear relationship. We can also extend these questions and concepts to relationships between groups of variables.

The relationships are only completely determined if we define a model for the joint distribution of all variables. The generalization of the **normal distribution** to the case of multiple variables forms the model that plays an even more central a role in multivariate statistics than in univariate.

c • In the iris plants data set, three groups are given by the three species. A first question is: **Are there differences between groups of observations?** As in univariate statistics, for this question we seek a statistical **test** that checks the null hypothesis of **equal expected values** (or equal values of another location parameter) or completely equal distributions for the groups.

d • The differences between groups in the examples of the iris plants and the fossils is primarily useful if they make it possible to classify new individuals whose group membership is unknown. We therefore ask: **To which group should a new individual** with unknown group membership be **assigned** on the basis of its features? This raises the question: **How sure can we be about such an assignment?**

Together, these questions form the starting point of **discriminant analysis**. Classification is often more accurately called **identification analysis** .

e • For the bank customers, there are no groups given, but we ask: **Can the observed individuals reasonably be divided into groups?** This is a fairly vaguely posed problem, and accordingly the field of **cluster analysis** allows for a variety of methods that produce groups with different properties.

f • A reasonable grouping requires that we be able to quantify the similarity or dissimilarity between observations. We thus ask: **How should the differences of values of multiple variables for two observed entities be summarized to one similarity or dissimilarity measure?** It can also be reasonable to ask about and similarity or dissimilarity between variables.

If we have determined dissimilarities between observations, then we can use them not only for creating groups, but also as the basis for additional **distance based methods** that mostly are used for graphical representation.

g • Two dimensional data can easily be represented in a scatter plot. We therefore ask: **Can we reduce multivariate data to two dimensions in a way that preserves the essential information?** Or less restrictively: **Can we reduce the dimension of the data without losing essential information?** In the spectra example, this question has a clear basis: Since, in the course of the reaction, only the concentrations of the few involved chemical substances change, the "significant differences" in the spectra must be captured with the changes of the concentrations. The desired dimension will therefore be the at most equal to the number of chemical substances.

One method for finding such dimensions in the sense of a descriptive method is **principal component analysis**. Models that formalize the idea are known as **factor analysis** and **linear mixing models**.

h • As simple analysis of variance in univariate statistics, the situation with multiple groups forms the starting point for the general models of analysis of variance and linear

regression. Analogously to these methodologies we ask: **How does the joint distribution of multiple variables of interest depend on (multiple) explanatory variables?** Again, the explanatory variables can be continuous variables or factors (nominal, categorial variables). Important goals of such an analysis are again inferences about causality or the prediction of the variable of interest from the explanatory variables.

Multivariate regression and **multivariate analysis of variance** (MANOVA) generalize the concepts and methods of the "usual" univariate regression and analysis of variance. (In some textbooks, the term "multivariate regression" is also used for regression with a single variable of interest and multiple explanatory variables, i.e. for *multiple* regression.)

i Most issues in multivariate statistics are therefore the same as those in univariate statistic, but concern **multiple variables of interest** simultaneously. New issues are those of cluster analysis and dimension reduction. However, the consideration of the joint distribution of multiple related variables of interest brings new concepts and ideas into play also in the methods generalized from univariate ones.

j **Data Mining.** In business, a key word has established itself: "data mining". It has to do with the analysis of data that, in commerce, accumulates in computer-legible form, primarily data about customers, stock, orders, and other transactions. In comparison to studies or experiments, this generally gives rise to huge data sets. However, what they contain is predetermined and usually can not be controled based on statistical issues. Often, the data must first be taken from different data bases and collected in a "**data warehouse**" before it is suitable for analysis.

The term "data mining" means to say that, with appropriate tools, precious resources are to be found in such a "mountain of data" . These appropriate tools include statistical procedures – besides the various regression methods, also the procedures from multivariate statistics. There are also algorithms that have been invented by computer scientists and engineers, which are often judged critically be statisticians.

The issues that exist in the data mining application area are

- The designation of all customers with certain features (database querying and management),
- Creating an overview of the customer data (description),
- The classification of the customers into certain groups (discriminant analysis),
- The search for possible groupings (cluster analysis),
- The prediction of a variables of interest, such as the future turnover of individual customers, from known explanatory variables (regression).

k **Categorial and Continuous Data.** Categorial variables are those that can take only finitely many possible values and thus indicate membership in a "category". In contrast, continuous variables can, in principal, take on any number as their values – sometimes limited to all non-negative numbers or to an interval. They are interpreted quantitatively. For a third type of variable, ordered discrete variables, only "discrete" values are possible, for example the whole numbers. These values have a reasonable order and may also be interpreted quantitatively.

In our examples, continuous data have been considered – except for the grouping variables, which are naturally categorial. When we talk about multivariate statistics, we usually think about continuous variables, which will also be true in this script.

Actually, all the problems presented can also be formulated for categorial features. We find little about this in books about multivariate statistics. An exception is Fahrmeir et al. (1996). A keyword in this area is "log-linear models", and this is often treated in books about categorial data and general linear models.

Ordered discrete variables can often be treated as continuous variables. Little is known about specific methods for such variables.

l **Relationship with Other Areas in Statistics.** For inferential statistics, we need probability models. The underlying model is – as expected – the normal distribution, here the multivariate normal distribution (3.2). It is also used in other areas of statistics:

- In the areas of **time series** and **spatial statistics**, initially only one variable that varies in time or space is considered. For the joint distribution of values for different time points or locations, normal distributions with special structure are suitable. If multiple variables are considered in time or in relationship to location, then in addition to using the multivariate normal distribution, a model in the sense of multivariate statistics must be considered (not done in this script).
- **Analysis of Variance models** with random effects for a single variable of interest can also be replaced by a model for the joint distribution of the values of the variable of interest for all experimental conditions.
- The multivariate normal distribution occurs as the **distribution of estimators** in all areas of statistics. Often, estimations for multiple parameters are calculated from the observations. We combine them into a "multidimensional statistic". Under the assumption of a one dimensional normal distribution for the observations, linear multidimensional statistics that are calculated from then are jointly multivariate normally distributed.

The multivariate variant of the central limit theorem says that many multidimensional statistics (even nonlinear) are approximately normally distributed with increasing number of observations – and this is true not only for normally distributed observations.

m **Geometry.** The concept of the joint distribution of two variables immediately calls to mind graphical representations like the scatter plot (1.2.b). We identify observations with points in the plane. This idea can be extended to three dimensions; from four it becomes more difficult and we hope that the insights that arise from the geometry of the two and three dimensional space also are valid for higher dimensions. Formally, the concepts of geometry can be formulated for the m dimensional space. The analogy to the concepts of multivariate statistics goes further and will aid comprehension. However, it appears that in higher dimensions, undesirable properties arise that contradict the two and three dimensional viewpoints. We speak of the "**curse of dimensionality**".

1.4 Software

a **Statistics Packages.** Most methods discussed in this script are available in comprehensive statistics software packages and can be used via the usual "graphical user interfaces".

The big classical statistics programs are called SAS and SPSS. They contain the most well-known methods in proven reliable form. The program S-Plus and the free software version R are in some ways less mature, but are based on the S language, to which we will return. Further programs that are generally less comprehensive but cheaper are Systat, Stata, ...

In this script, the S language functions that implement the presented methods are summarized in a section at the end of a chapter.

b **Data Mining.** Because of a strong market, the large software houses offer modules that are specially designed for applications in data mining. For such modules, it is characteristic that

- they can operate well with huge databases,
- they are simple to use,
- besides the most important classical procedures, they contain some ad-hoc procedures (i.e., algorithms) with good marketing.

The most famous packages are "Clementine", which is associated with SPSS, and "SAS Data Miner".

c Since the analysis of larger empirical studies can become fairly complex, the use of a command oriented statistical software has proved itself, in other words a **user-friendly programming language** with a comprehensive collection of statistical functions.

Essential to such a language is that it

- is vector and matrix oriented and
- the results of all functions are stored in some form that makes their further use easy.

d **Statistics Language S.** One such language is called "S". There is a commercial implementation by the name **S-Plus** and a free software variant named "**R**". The two variants are largely identical in syntax and the available basic functions.

The S language has become a means of communication for statisticians the world over. New procedures are made available as (free) "libraries" by the researchers and can be used without much additional work.

This large flexibility and comprehensiveness means that not all parts of the system are equally reliable and error notifications are often difficult to interpret.

e Besides S, the software packages Matlab and Mathematica fulfill the previously mentioned conditions of matrix orientation and results storage. However, their statistics functions are much less comprehensive and, in some part, ill designed.

f As introductions to the S language, there are books from **?**), **?**), **?**), **?**). Guides of varying lengths can be found on the web at `www.R-project.org`.

1.5 About This Script

a **Objective.** This script is primarily conceived for the postgraduate course in applied statistics. In this script, problems, and the reasonable application of their corresponding methods, stand in the foreground. Formal mathematical theory is largely absent. Still, linear algebra is introduced and used, since it is very helpful for understanding the concepts and models.

b **Assumptions.** This script assumes that the basic concepts of (frequentist) univariate statistics are familiar. Basic knowledge of matrix calculations (or linear algebra) are advantageous; they are repeated briefly in chapter 2 .

c **Educational Objectives**

- For applied problems, you should be able to identify the appropriate methods.
- You should be familiar with the basic ideas that underlie the classical methods in multivariate statistics.
- You should be able to estimate the potential of some methods of explorative multivariate analysis.
- You should understand the models on which these procedures are based and therefore be able to evaluate the goals, assumptions, and limitations of the methods.

d **Order.** According to experience, the chapter about the models can be a hurdle for application oriented learners, which can have a demotivating effect after the repetition and application of linear algebra. Thus, in the lecture I insert the sections about principal components and biplot (7.1 - 7.2) between them.

e **Notation** In this script, brackets are used in an unorthodox, but considered, way:

- $\{..\}$ The curly brackets are only used for sets.
- $\langle..\rangle$ These angular brackets enclose arguments of functions.
- $(..)$ The plain parentheses show the priority of calculation operations (as in $(a+b)c$).
- $[..]$ The square brackets are used for vectors and matrices.

L Literature

a There are many books about multivariate statistics. They are more comprehensive and usually more demanding than this script. The following books are especially recommended:
Mardia, Kent and Bibby (1979): Comprehensive. Theory and examples. Very good.
Rencher (1995): Roughly corresponds to this script in level. Very understandable.
Rencher (1998): Reference work with many tips about literature (without cluster analysis and distance methods).
Krzanowski (1988): Applied. Fahrmeir et al. (1996): German, very good, unfortunately very expensive.

The book by Flury (1997) is very clear, applied, and didactic, so very readable. It includes only a part of the lecture and is still quite thick.

The book by Gnanadesikan (1997) has explorative data analysis as its focus and set milestones at its first appearance in 1977.

b **More textbooks** about multivariate statistics: Anderson (1984), Morrison (1967, 1976), Muirhead (1982), Bilodeau and Brenner (1999), Seber (1984), Srivastava and Carter (1983), Chatfield and Collins (1980), Cooley and Lohnes (1971), Green and Carroll (1976), Harris (1975), Johnson and Wichern (1982, 1988), Karson (1982), Kendall (1957, 1961), Manly (1986, 1990), **?**)

c **Specialities.** Maxwell (1977) (behavioral science) and Tatsuoka (1971) (psychology) write for special application areas.

Special Themes: Distributions: Johnson and Kotz (1972);
Models for Measurement Error: Fuller (1987), Brown (1993);
Missing Data: Schafer (1997), Little and Rubin (1987);
Graphics: Everitt (1978).