# 15

# INTERPRETING SEVERAL TYPES OF MEASUREMENTS IN BIOSCIENCE

Achim Kohler

*Matforsk, Ås, Norway; and Norwegian University of Life Sciences, Ås, Norway*

Mohamed Hanafi, Dominique Bertrand, and El Mostafa Qannari

*ENITIAA, NANTES, France*

Astrid Oust Janbu

*Norwegian Water Technology Centre, Oslo, Norway*

Trond Møretrø and Kristine Naterstad

*Matforsk, Ås, Norway*

Harald Martens

*Matforsk, Ås, Norway; Norwegian University of Life Sciences,*
*Ås, Norway; and University of Copenhagen, Frederiksberg, Denmark*

## 15.1 INTRODUCTION TO THE ANALYSIS OF SEVERAL DATA SETS

Due to the rapid development of high-throughput instruments in biomedicine during recent years, a mismatch between the large amount of data produced and the capacity to analyze these data has arisen. Often single instruments are producing several hundreds up to several hundred thousands of variables of a given type for each sample. The rapid instrumentation development also allows collecting different types of variables from a given set of samples.

*Biomedical Vibrational Spectroscopy,* Edited by Peter Lasch and Janina Kneipp

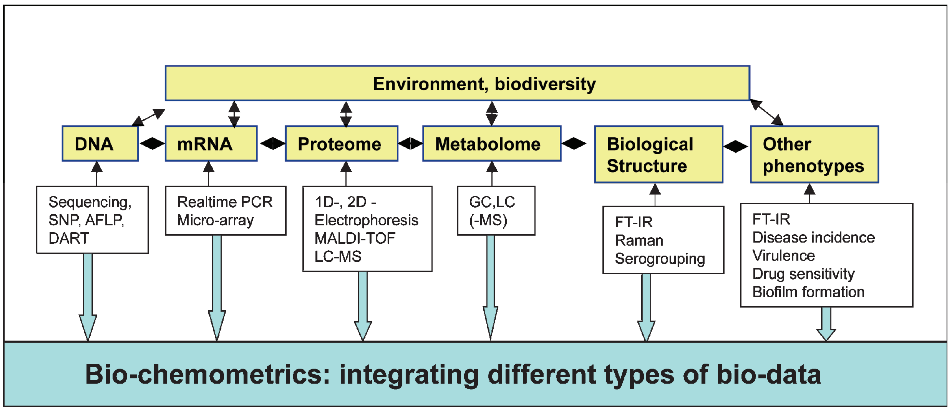
Environment, biodiversity
DNA
mRNA
Proteome
Metabolome
Biological Structure
Other phenotypes
Sequencing, SNP, AFLP, DART
Realtime PCR Micro-array
1D-, 2D - Electrophoresis MALDI-TOF LC-MS
GC,LC (-MS)
FT-IR Raman Serogrouping
FT-IR Disease incidence Virulence Drug sensitivity Biofilm formation
Bio-chemometrics: integrating different types of bio-data

Figure 15.1. Illustration of how in systems biology data are obtained along the causal chain from genotype to phenotype. The obtained data are typical multiblock data; that is, different measurement principles are applied to the same *N* samples resulting in many multivariate data blocks.

This is illustrated in Fig. 15.1 for functional genomics: Biological systems can nowadays be studied by high-throughput instrumentation at all levels along the causal chain from genotype to phenotype. Examples for the different methodologies are as follows: (a) sequencing or single nucleotide polymorphism (SNP) technologies to study DNA diversity, (b) micro-array techniques to study gene expression, (c) mass-spectroscopic MALDI-TOF or electrophoresis to obtain proteome data, and (d) different biospectroscopic and chromatographic techniques to study the metabolome, other phenotype characteristics such as drug resistance, biological structure and so on, as well as available background information about the samples at hand.

The use of vibrational biospectroscopic techniques for these studies will gain in importance due to the ease of their use, their high specificity for latent phenotype information, and consequently their high potential for screening. When all these different levels of information (or at least some of them) are obtained for the same samples, covariations and interactions between the different stages along the causal chain can be studied, leading to in-depth insight into the underlying patterns of variation and mechanisms of control. In the future the amount of collected data will be so large that professional databases will be needed to deal with it. However, there is also a need for adequate data analysis tools to link the different data sets.

Since most biological systems are very complex, and since many modern measurement techniques yield multichannel data, multivariate data modeling is useful. This chapter outlines some relatively new multivariate methods that help the researcher to handle multiblock situations. A typical multiblock situation is illustrated in Fig. 15.2: Data from several blocks $b = 1, \ldots, B$ (different measurement types and/or subsets of instrument channels from a given type) have been obtained from the same set of samples, $i = 1, \ldots, N$. Within each block $b$ a certain set of variables $k = 1, \ldots, K_b$ has been measured, resulting in an $N \times K_b$ data table (matrix block) $\boldsymbol{X}^{\mathrm{b}}$, $b = 1, \ldots, B$. All measurements are performed for the same $N$ samples. By sorting the $N$ samples in the same order in every block, the necessary row correspondence is achieved.

The present chapter focuses on one family of multivariate data modeling methods, based on so-called bilinear modeling. These provide data-driven (rather then theory-driven) mathematical descriptions of the main content of whole data tables. When the data from only one single block are to be screened, the one-block technique of principal component

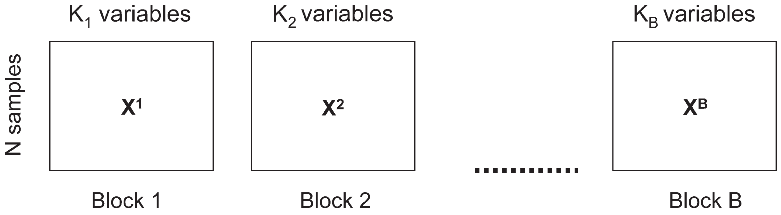

Figure 15.2. Illustration of a typical multiblock situation. Every block (square) is representing one data matrix, where rows are referring to samples and columns to variables. The number of rows *N* is the same for every data block, and the number of columns $K_1, \ldots, K_B$ is in general different for every block.

analysis (PCA) is useful. For relating two different blocks to each other (e.g., for quantitative multivariate calibration), the technique of partial least squares regression (PLSR) is a versatile tool. When extended to the simultaneous analysis of several blocks $\boldsymbol{X},^b\ b = 1, \ldots, B$ as presented in Fig. 15.2, it is called a multiblock analysis. The scope of the multiblock analysis depends on the application, but in biospectroscopy the following general objectives may be stated: (a) to find the main underlying patterns of covariation that are common for several or all of the blocks, (b) to measure how strongly these covariant patterns are represented in each block and how they relate the blocks predictively to each other, and (c) to visualize how already known bands contribute to the covariance patterns.

The mathematical modeling philosophy here is "soft" – more or less purely data-driven: Contrary to more classical mechanistic mathematical modeling, a minimum of causal beliefs are imposed. Contrary to more classical statistical modeling, a minimum of distributional assumptions and independence requirements are imposed. The idea is to "let the data talk for themselves."

But the user's background knowledge is still important – in the planning of experiments, in the preprocessing of data, and in the interpretation of the results. Hence, while this chapter outlines some mathematical data modeling methods, the reader should also note the importance of nonmathematical aspects – in particular, graphical inspection of the input data – of intermediate steps and of the final output. Data-driven modeling can only work if the data are informative. Therefore, it is essential to plan the experiments well, so as to ensure sufficient variations among the samples, along with sufficient precision and relevance among the variables.[1]

Before mathematical modeling, the raw input data should be inspected, at least superficially, for gross errors. During the mathematical modeling the statistical results should be inspected critically in light of background knowledge. After the modeling, the conclusions should be checked, for quality assurance of the data analytical process and for easy communication to others, by plotting those input variables that the modeling has shown to be most important.

Data preprocessing is an important mathematical aspect of data modeling that will not be treated much in this chapter. For example, measured transmission spectra T are usually filtered to reduce noise and are transformed into, for example, absorbance = log(1/T) to ensure linear response to chemical sample composition according to Beer's Law. Moreover, it is often useful to distinguish physical effects from chemical effects during preprocessing – for example, separating multiplicative light scattering and sample thickness from additive baseline effects and chemical light absorbances. This can be attained, for example, by taking spectral derivatives, followed by so-called standard normal variate (SNV) filtering or the more model-based preprocessing, the extended multiplicative signal correction (EMSC).[2–5]

The simplest multivariate analysis is to perform PCA on every single block $\boldsymbol{X}^{b}$. This approach will only reveal structures within each block; it will not in general reveal common structures between the blocks. However, by visually comparing score plots for the separate PCA block models, the researcher may identify some common variation pattern visually. In order to develop systematically the logic and algorithms of the multiblock analysis and the tools for visualization, we will start in Section 15.2 by introducing the nonlinear iterative partial least-squares (NIPALS) algorithm for PCA. This is historically how the originally psychometric PCA method was introduced into chemometrics.[6] Once this simple algorithm is understood – in terms of how the model elements (scores, loadings, correlation loadings and residuals) are obtained – it is easier to understand the plots that we will use extensively for finding correlated interpretable signals in the different blocks.

In Section 15.3 we will introduce the PLSR for the analysis of two data blocks. This will bridge to the multiblock method "multiblock principal component analysis" that will be introduced in Section 15.4. In Section 15.5 we will give a short overview over other multiblock approaches.

Sections 15.2–15.3 are organized in the following way: Every section starts by introducing the respective method. The process of data analysis and the NIPALS algorithm is then presented schematically in a frame, followed by an example illustrating this process. A reader who does not want to go into mathematical details may ignore the frames.

*Example.* An example from microbiology will be used for illustration: *Listeria monocytogenes* is a food-borne bacterium that may cause serious infections.[7] Most outbreaks of *L. monocytogenes* have been associated with foods that are stored at low temperature and are consumed without heat treatment. Since the market for ready-to-eat products is increasing, this bacterium has been in focus during the recent years. It has been suggested to control the growth of *L. monocytogenes* in food by adding antimicrobial peptides called bacteriocins.[8,9] Such peptides are produced by lactic acid bacteria and are generally recognized as safe to consume. The mechanism of action of bacteriocins is being investigated, and it is yet not fully understood for any of the different classes of bacteriocins. Studies point toward the presence of a specific molecule in the cell membrane of the target bacterium necessary for the bacteriocin to perform its activity (pore forming and dissipation of the proton motive force).[10–12] Other influencing factors on the mechanism behind *L. monocytogenes* susceptibility toward bacteriocins seem to be membrane fluidity, cell wall density, and charge of membrane or cell wall.[13] Knowledge of the mechanism of action is important to be able to elucidate and thereby prevent development of resistance toward such an agent.

In a previous study, 200 different *L. monocytogenes* strains were exposed to the bacteriocin sakacin P, and it was found that the susceptibility toward this bacteriocin varied between the strains (determined by using a bioassay with *Listeria ivanovii* as indicator and purified sakacin P as standard). Surprisingly, the strains formed two groups according to their susceptibility level, separated by a gap when sorted as shown in Fig. 15.3.[14]

Such a gap has not been reported for other bacteriocins, and the biochemical and genetic basis behind this grouping of *L. monocytogenes* is not known. Other studies – also related to properties of the cell wall/cell membrane, investigating zero-type and the ability to grow in the presence of an antibacterial agent, benzalkonium chloride (BC), and the bacteriocin nisin – have been performed and altogether have generated large data sets for these strains.

More generic analyses – as genetic fingerprinting [amplified fragment-length polymorphism (AFLP)], spectroscopic fingerprinting (FT-IR and RAMAN), and protein profiling (SDS-PAGE) – have also been performed.[15–17] For AFLP and FT-IR analysis a subset of 88 strains was analyzed: The 20 strains closest to the sakacin susceptibility gap

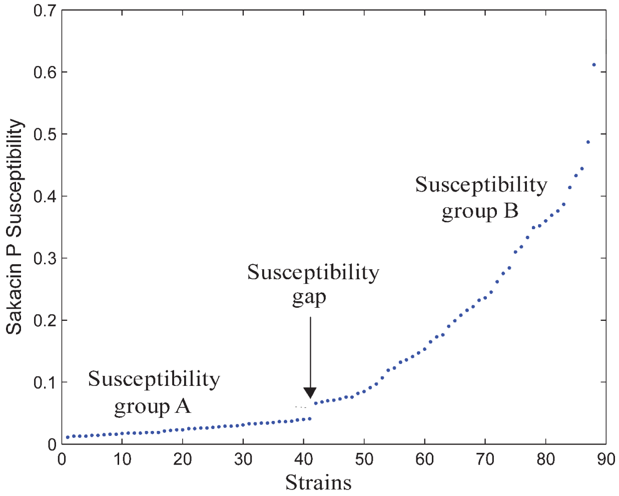

Figure 15.3. Susceptibility level of the 88 *Listeria monocytogens* strains according to Ref. 14. The strains are sorted with respect to susceptibility.

(10 on each side) and 68 strains covering the whole susceptibility range were chosen (including the 20 most extreme samples with respect to susceptibility – that is, the 10 with lowest and the 10 with highest susceptibility). By the collected data three different types of data according to Fig. 15.1 are represented: “DNA” (AFLP), “Biological Structure” (FT-IR, Raman, SDS-PAGE, sero grouping) and “Other Phenotypes” [susceptibility to sakacin P, nisin, and the antibacterial agent benzalkonium chloride (BC)]. Integrating and exploiting of these data sets in combination will give further information about the background behind the *L. monocytogenes* strain to strain variation in general and especially the variation in susceptibility to bacteriocins. The latter may also lead to further understanding of the mechanism of action of bacteriocins. In this chapter, AFLP data (DNA), FT-IR data, and serotyping (biological structure) and susceptibility to sakacin P, nisin, and benzalkonium chloride (other phenotypes) will be used to illustrate the methods presented for the analysis of several data sets simultaneously.

## 15.2 PRINCIPAL COMPONENT ANALYSIS OF ONE DATA TABLE

By PCA the user can get an overview of the main patterns of covariation within a data table. PCA may be regarded as “the mother of all multivariate modeling methods,” because it is so simple and so powerful.

For illustration of PCA, consider an input data table consisting of $N$ (e.g., 100) samples (rows, “objects,” e.g., bacterial strains), chosen as interesting and expected to be related in some way, and $K$ (e.g., 2000) input variables (columns, “descriptors,” e.g., wavenumber channels), chosen as possibly informative for these samples, and scaled so that they are expected to have approximately the same noise level: The $N \times K$ data Table $\mathbf{X} = [x_1, x_\mathbf{2}, \ldots,$ $\boldsymbol{x_k}, \ldots \boldsymbol{x_K}]$ with columns $\mathbf{x_k} = [x_{1,k}, x_{2,k}, \ldots, x_{\mathrm{i},k}, \ldots, x_{N,k}]'$, for example, $N = 100$ samples (rows) and $K = 2000$ input variables (descriptors, e.g., wavenumber channels) – is approximated by a mean profile $\bar{\mathbf{x}}$ (symbolized by the bar above $\mathbf{x}$) plus a mean-centered table of variations around this mean: $\mathbf{X} = \bar{\mathbf{x}} + \mathbf{X}_0$. Systematic patterns of covariations in $\mathbf{X}_0$ are now to be searched for in terms of so-called principal components (PCs). The process is

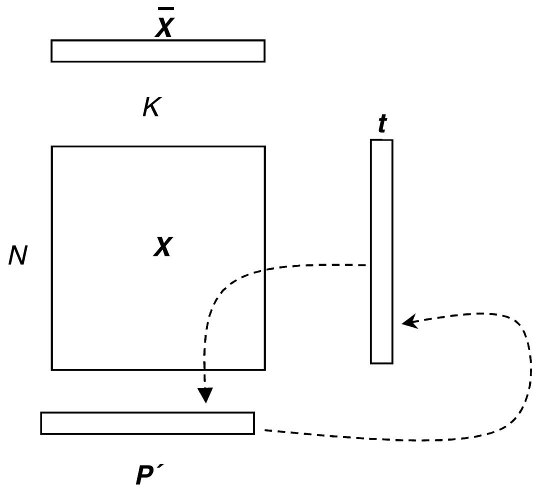

Figure 15.4. Illustration of the NIPALS algorithm (see Frame 1) and the relation between scores **t** and loadings **p** for PCA.

illustrated in Fig. 15.4; for simplicity, only one such PC is shown, along with arrows that outline how it is obtained from the data.

Each PC is represented by a column vector **t** and a row vector $\mathbf{p}'$. The score vector $\mathbf{t} = [t_1, t_2, \ldots, t_N]'$ is a “supervariable” whose “concentrations” in the $N$ samples is a linear combination (i.e., a weighted sum) of the mean-centered $X$-variables. The transposed loading vector $\mathbf{p}' = [p_1, p_2, \ldots, p_K]$ characterizes this “supervariable” in terms of the $K$ individual input variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\mathrm{K}$. Loading vector **p** may be thought of as the average difference spectrum between those samples that have high levels **t** and those that have low levels **t**. The vector pair **t** and **p** is found by a search process (see below). The outer vector product of the score and loading vector pair, $\mathbf{tp}'$, defines the contribution of this PC to the modeling of data table **X**. Hence, using only one PC, the input data **X** are approximated by the rank-one “bilinear model” $\mathbf{X} = \bar{\mathbf{x}} + \mathbf{t}_1\mathbf{p}_1' + \mathbf{X}_{(1)}$, where $\mathbf{X}_{(1)}$ is the residual after one PC. If the data table **X** contains $A$ different types of systematic variation patterns, these are picked up by a sequence of $A$ PCs numbered $a = 1, 2, \ldots, A$, each with less importance than the previous ones, plus the unmodeled residual:

$$\mathbf{X} = \bar{\mathbf{x}} + \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2 + \cdots + \mathbf{t}_a\mathbf{p}_\mathrm{a}' + \cdots + \mathbf{t}_A\mathbf{p}_A' + \mathbf{X}_{(A)}$$

Several equivalent algorithms exist for extracting the PCA scores and loadings. The NIPALS estimation algorithm is illustrated by the arrows in Fig. 15.4. “**X**” in the figure symbolizes the information in the $N \times K$ input data table that remains after extracting the $a-1$ previous PCs, $\mathbf{X}_{(\mathrm{a}-1)} = \mathbf{X} - \bar{\mathbf{x}} - \mathbf{t}_1\mathbf{p}_1' - \mathbf{t}_2\mathbf{p}_2 - \cdots - \mathbf{t}_{a-1}\mathbf{p}'_{\mathrm{a}-1}$. This is the variation available for finding **t** and **p** for the next PC, #$a$. Hence, for the first PC ($a = 1$), “**X**” represents the mean-centered variables “**X**” $= \mathbf{X}_{(0)} = \mathbf{X} - \bar{\mathbf{x}}$. The NIPALS search algorithm is started by choosing an arbitrary score vector **t** – for example, $N$ random numbers, or (more preferably) the column in **X** with highest remaining variation. The $K$ columns in data matrix **X** are projected on the score vector **t** in order to calculate the $K$ elements in loading vector **p**. In other words, from the approximation model for each input variable, $\mathbf{x}_k \approx \mathbf{t}p_k'$, each column k in the matrix **X** is regressed “vertically” on **t** by least-squares regression to estimate the regression coefficient $\mathrm{p_k}$, for $k = 1, 2, \ldots, K$. To stabilize the iterative procedure, the loading vector **p** is normalized to have a sum of squares of 1. Then, from the approximation model for each input sample, $\mathbf{x}_i \approx \mathbf{t}_i\mathbf{p}'$, each of the $N$ rows in matrix **X** is conversely regressed “horizontally” on $\mathbf{p}'$ in order to obtain an updated estimate of the $N$

elements in score vector $\mathbf{t}$. This iterative search procedure is performed until $\mathbf{t}$ and $\mathbf{p}$ change no more.

It has been shown that the NIPALS algorithm for extracting the PCA component #$a$ is equivalent to finding the main variation in $\mathbf{X}$ by diagonalizing the covariance matrix of "$\mathbf{X}$" $= \mathbf{X}_{(a-1)}$. Consequently, the loading vector $\mathbf{p}$, found by the NIPALS algorithm of the matrix "$\mathbf{X}$," refers to the first principal component of the residual data matrix $\mathbf{X}_{(a-1)}$. In other words, the loading vector $\mathbf{p}$ defines the direction in the variable space that maximizes the variation of all objects of $\mathbf{X}$. The score vector $\mathbf{t}$ shows the realization of this PC in the individual samples. Once the first PC has been estimated, the effect of its parameters $\mathbf{t}_1 = \mathbf{t}$, $\mathbf{p}_1 = \mathbf{p}$ is subtracted: $\mathbf{X}_{(1)} = \mathbf{X}_{(0)} - \mathbf{t}_1\mathbf{p}_1^{'}$. The second principal component is then calculated by the NIPALS algorithm applied to the residual matrix "$\mathbf{X}$" $= \mathbf{X}^{(1)}$, yielding $\mathbf{t}_2 = \mathbf{t}$, $\mathbf{p}_2 = \mathbf{p}$. From $\mathbf{X}_{(2)} = \mathbf{X}_{(1)} - \mathbf{t}_2\mathbf{p}_2^{'}$, the residual "$\mathbf{X}$" $= \mathbf{X}_{(2)}$ is defined and the third PC is searched for, and so on.

The calculation of the residual by subtracting the effect of the consecutive components is called deflation. For PCA there is only one possibility to deflate (described above). For the analysis of two data blocks by partial least-squares regression or several data blocks by multiblock PCA, there are several possibilities for deflation. The deflation procedure is a very important step since it defines orthogonality properties of loadings and scores. In PCA both score vectors and loading vectors are orthogonal, leading to the very special situation that the components are independent in both variable space and sample space. Mathematically, each consecutive PCA component $\mathbf{t}_a$, $a = 1, 2, \ldots, A$, is defined as the first eigenvector of the residual covariance between the variables in $\mathbf{X}$.

In the following, index $a = 1, 2, \ldots, A$ represents PC #, just like $k = 1, 2, \ldots, K$ represents input variable # and $i = 1, 2, \ldots, N$ represents sample #. After $A$ estimation and deflation steps, $A$ principal components have been obtained. The score vectors for the $A$ components can be summarized by the score matrix $\mathbf{T}_A = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_a, \ldots, \mathbf{t}_A]$, and equivalently their loading vectors can be summarized by the matrix of loadings $\mathbf{P}_A = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_a, \ldots, \mathbf{p}_A]$.

A versatile concept that will be used throughout this chapter for visualization of the main bilinear structures found (e.g., Fig. 15.5b) is the so-called correlation loading plot. In the correlation loading plot the correlations between variables and scores are plotted. More precisely, let $\mathbf{x}_k$ be a variable ($k$th column of $\mathbf{X}$) and $\mathbf{t}_m$, $\mathbf{t}_n$ the score vectors of two PCs, say $m = 1$ and $n = 2$. For the correlation plot, the correlation coefficients $\mathrm{r}_{km}$ and $\mathrm{r}_{kn}$ relating $\mathbf{x}_k$ to $\mathbf{t}_m$ and $\mathbf{t}_n$ are computed, and the variable $\mathbf{x}_k$ is represented as a point with coordinates $[\mathrm{r}_{km}, \mathrm{r}_{kn}]$. This is done for every variable $k = 1, 2, \ldots, K$ in $\mathbf{X}$. But if there are many variables, the plot may become "crowded," and only the most salient variables are named explicitly. An important point is that, due to the orthogonality between scores, the norm of the vector $[r_{km}, r_{kn}]$ is by necessity less or equal to 1. It is thus useful to draw, on the correlation plot, a circle of radius equal to 1, with the center at the origin of the graph. A position near the circle in this plot indicates that it is possible to exactly predict $\mathbf{x}_k$ from the scores $\mathbf{t}_m$, $\mathbf{t}_n$. On the other hand, a variable positioned close to the origin of the graph is not predictable from the studied pair of scores, and it plays almost no role in the "construction" of these scores. Moreover, if two variables $\mathbf{x}_k$, $\mathbf{x}_j$ are close together on the correlation plot for $\mathbf{t}_1$, $\mathbf{t}_2$, and also close to the perimeter of the circle, they are strongly positively correlated, while they are strongly negatively correlated if they are found at the opposite perimeter.

*Example.* The score plot of the principal component analysis of the FT-IR spectra of the set of 88 *Listeria monocytogenes* strains ($N = 88$) as described above and in Ref. [15] is

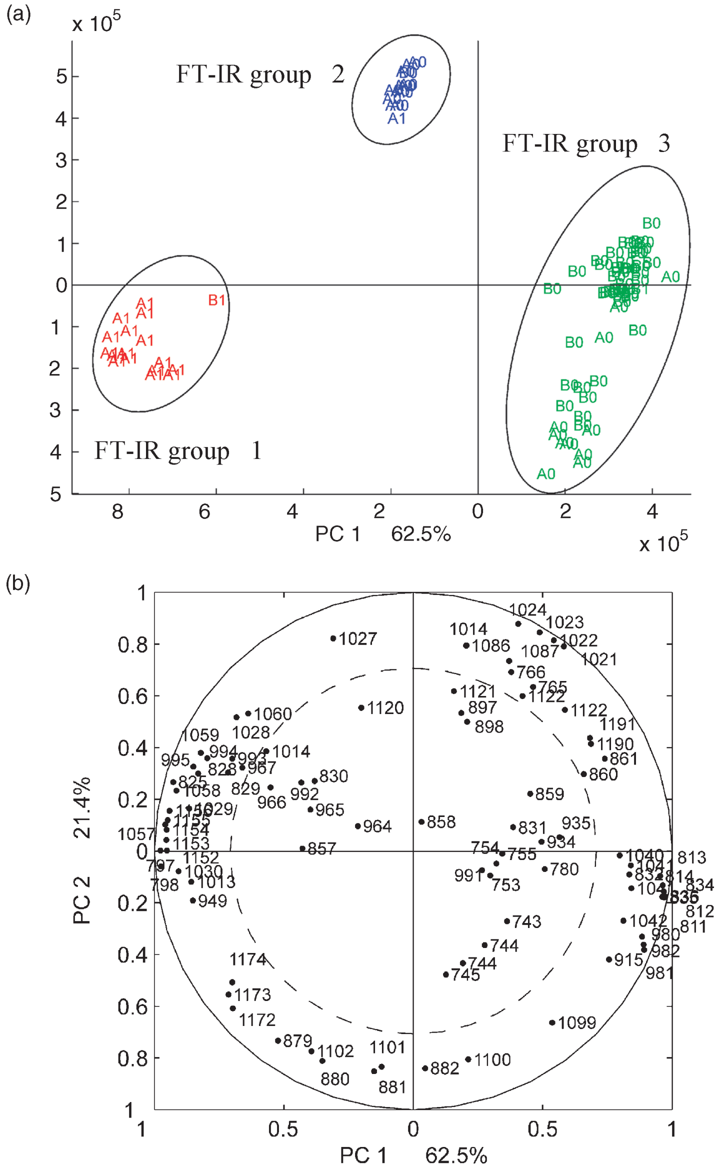

Figure 15.5. Principal component analysis of the FTIR data in the spectral range 720–1200 cm$^{-1}$. (**a**) Score plot with symbols A and B representing the sakacin P group, numbers 0 and 1 marking serotype $^1/_2$ and 4, respectively, and colors coding the three visible FTIR groups (1, red: 2, blue: 3, green). (**b**) The corresponding correlation loading plot.

shown in Fig. 15.5a for the spectral range from 720–1200 cm$^{-1}$. This region includes mainly bands from polysaccharides and backbone vibrations.

Before PCA the spectra were preprocessed to correct for unwanted physical baseline and scaling variations by extended multiplicative signal correction,[2,18] and then mean-centered as described in Frame 1. The score plot visualizes the main variation in the data set; the origin (0,0) corresponds to the "average" sample, with spectrum $\bar{\mathbf{x}}$. The first principal component accounts for the largest variation, while the second component

**FRAME 1**
**Analysis of a Single Data Table by Principal Component Analysis (PCA)**

The PCA Process

- Select a set of samples and a set of variables.
- Scale each variable in order to obtain approximately compatible variation ranges, and form data table **X** (rows = samples, columns = variables).
- Compute and subtract the mean sample: $\mathbf{X}_0 = \mathbf{X} - \bar{\mathbf{x}}$.
- Model data table $\mathbf{X}_0$ as sum of a few trustworthy principal components (PCs) $a = 1,2,\ldots,A$ plus residual after $A$ PCs, $\mathbf{X}_{(A)}$:

$$\mathbf{X}_0 = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \cdots + \mathbf{t}_a\mathbf{p}_a' + \cdots + \mathbf{t}_A\mathbf{p}_A' + \mathbf{X}_{(A)}$$

that is,

$$\mathbf{X} = \bar{\mathbf{x}} + \mathbf{T}_A\mathbf{P}_A' + \mathbf{X}_{(A)}$$

- First, compute a higher number of PCs (e.g., 10) than necessary. Then determine the optimal number of PCs to trust, $A$ (e.g., 3), by cross-validation, noise assessments, and so on, and leave the remaining PCs $A+1, A+2, \ldots$ as unmodeled residual $\mathbf{X}_{(A)}$. Study the $A$ PCs graphically: Look for covariation patterns, clusters, and so on. Inspect residual $\mathbf{X}_{(A)}$ for possible outliers, and check the effect of reanalyzing without them.

The parameters **t** and **p** can be estimated for one PC at a time—for example, by the NIPALS algorithm. For component #$a$, the NIPALS algorithm searches for a new PC to approximate the residuals after the previous $a - 1$ PCs, "**X**" = $\mathbf{X}_{(a-1)}$, by the following steps:

- Choose a start score **t**.
- Iterate the following steps until convergence:
  - Loading vector $\boldsymbol{p} = \frac{\mathbf{X}'\mathbf{t}}{\mathbf{t}'\mathbf{t}}$.
  - Normalize **p** to length 1 by $\mathbf{p} = \frac{\mathbf{p}}{\sqrt{\mathbf{p}'\mathbf{p}}}$.
  - Score vector $\mathbf{t} = \mathbf{X}\mathbf{p}$, or equivalently, $\mathbf{t} = \frac{\mathbf{X}\mathbf{p}}{\mathbf{p}'\mathbf{p}}$.
  - Check convergence—that is, if **t** and **p** no longer change.

accounts for the second largest variation, and so on. The first two principal components were found to account for the most part of the variation between the 88 samples in this spectral region (62.5% and 21.4%, respectively; total 83.9%). Figure 15.5a shows their sample score configuration ($\mathbf{t}_1$ versus $\mathbf{t}_2$). The samples are found to show three distinct clusters, here named FT-IR groups 1 (red), 2 (blue), and 3 (green), respectively. The samples are labeled according to their sakacin-susceptibility and serotype: (A) and (B) mean the sample has a susceptibility value below or above the gap in Fig. 15.3, respectively; (0) and (1) mean the samples have serotype $^1/_2$ and 4, respectively. Fig. 15.3 shows that the susceptibility of these strains is widely spread over a whole susceptibility range. The three FT-IR groups separate more or less clearly with respect to serotype, and to some extent also with respect to sakacin sensitivity.

In Fig. 15.5b the corresponding correlation loading plot for these first two PCA components is shown. Only wavenumbers that were referring to band maximum or -minimum positions in the FT-IR spectrum loadings are plotted. Since band positions may shift slightly from sample to sample, there are in general several wavenumber positions that are referring to the same chemical bond in Fig. 15.5b. The figure shows the correlation between variables and PCs. Variables that are close to one (outer circle) are well-explained by these first two PCs (near 100% variance explained). The inner circle corresponds to 50% explained variance. The two plots should be read together, with directions from the origin named like the handle directions of a watch: Horizontally along PC1, FT-IR group 3 in Fig. 15.5a is situated primarily at around "3 o'clock" (spread between 2 and 5 o'clock). Consequently, the corresponding *X* variables around "3 o'clock" in Fig. 15.5b (981, 982, . . . , 813 $\mathrm{cm}^{-1}$) are higher than average in FT-IR group 3, but lower than average in the other FT-IR groups. Conversely, *X* variables around "9 o'clock" in Fig. 15.5b (949, 798, . . . , 828 $\mathrm{cm}^{-1}$) are higher than average in group 1 and lower than average in group 3. At about "12 o'clock" the plots show that FT-IR group 2 has higher-than average absorbance (at 1027, . . .) and lower-than-average absorbances at the opposite perimeter, 879, 880, 881 882, 1100, 1101, 1102 $\mathrm{cm}^{-1}$.

The example has shown that PCA can be used to explore a single data table for dominant latent structures, that is, systematic patterns of covariation – and to relate both samples (bacteria strains) and variables (FT-IR wavenumbers) to these latent structures. The visual inspection of the loading and score plot allowed us to discover unexpected clustering of samples and to interpret their differences in terms of the variables. However, the way that background information was brought in was only qualitative (letters A and B, numbers 1 and 0 in Fig. 15.5a; wavenumbers in Fig. 15.5b). In order to relate, for example, FT-IR spectra and different susceptibility values to each other quantitatively, a two-block method must be used.

## 15.3 SIMULTANEOUS ANALYSIS OF TWO DATA BLOCKS BY PARTIAL LEAST-SQUARES REGRESSION (PLSR)

For the analysis of relationships between two data blocks $\mathbf{X}$ and $\mathbf{Y}$, many different methods may be used. The partial least squares regression is one useful method. It owes its versatility to a combination of two aspects – bilinear approximation and linear regression: On one hand, just like PCA, it extracts a few (*A*) bilinear latent variables or PLS components (PCs) $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_A]$ from the "regressor" matrix $\mathbf{X} : \mathbf{T} = f(\mathbf{X})$. These *X*-scores are used for modeling each variable in both $\mathbf{X}$ and $\mathbf{Y}$: $\mathbf{X} \approx \mathbf{T}_A \mathbf{P}'_A$, $\mathbf{Y} \approx \mathbf{T}_A \mathbf{Q}'_A$. On the other hand, just like traditional multiple linear regression (MLR), it employs a linear regression model with a predictive direction $\mathbf{Y} \approx \mathbf{X}\mathbf{B}_A$, but with regression parameters $\mathbf{B}_A$ estimated via the *A*-dimensional bilinear PLSR model. In analogy to the single-block PCA model, the two-block PLSR model with *A* components may be written $\mathbf{X} = \bar{\mathbf{x}} + \mathbf{T}_A \mathbf{p}_A^{'} + \mathbf{X}_{(A)}^{'}$, $\mathbf{Y} = \bar{\mathbf{y}} + \mathbf{T}_A \mathbf{Q}_{\mathrm{A}}^{'} + \mathbf{Y}_{(A)}$. In contrast to PCA of $\mathbf{X}$ alone, the MLR-like aspect of PLSR ensures that the PCs $\mathbf{T}_A = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_A]$ from "regressors" $\mathbf{X}$ are relevant also for "regressands" $\mathbf{Y}$. In contrast to traditional full-rank MLR, the low-rank PCA-like aspect of PLSR ensures that graphical overview is attained and that the rank-*A* regression model is statistically stabilized against measurement errors in block $\mathbf{X}$ and/or $\mathbf{Y}$, even for data sets with highly inter-correlated input variables such as biospectroscopy wavenumber channels. These aspects make PLSR useful for multivariate calibration.[19] In the present multiblock setting, the way these two aspects are combined algorithmically makes PLSR particularly informative because it facilitates graphical overview of the reliable and relevant relationships in rather complex data sets. This will be demonstrated below.

The PLSR model parameters may be estimated in several different, but equivalent ways. In simple words, the PLS components find and reveal the main covariations within

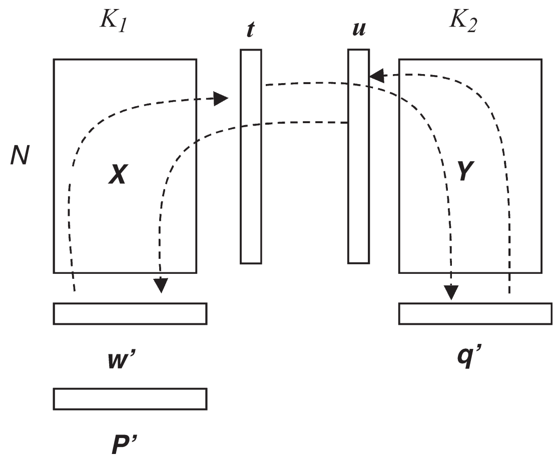

Figure 15.6. Illustration of the NIPALS algorithm (see Frame 2) and the relation between scores and loadings for PLSR.

and between blocks **X** and **Y**. Mathematically, each consecutive PLS component is defined by the first eigenvector of the residual covariance between **X** and **Y**.

The meaning of the PLSR model parameters may be easier to understand by seeing how the iterative NIPALS algorithm now works; this is outlined in Fig. 15.6 and described in Frame 2.

Just like in PCA, the NIPALS algorithm is employed to estimate the parameters (loadings and scores), for one PLSR PC at a time: The effect of an obtained PC is removed from the input data **X** and **Y** by deflation and the same process started again for the next PC. Each PC $\mathbf{t}_a$, $a = 1,2, \ldots, A$ is obtained from block **X** like in PCA, but now using block **Y** for guidance. The NIPALS algorithm seeks to approximate the covariation that remains after removal of the $a-1$ previous PCs, within and between the input matrices **X** and **Y**. Hence, input data for PC #*a* are the residual matrices after removal of the $a-1$ previous PCs, $\mathbf{X}_{(a-1)}$ and $\mathbf{Y}_{(a\text{ - }1)}$, named "**X**" and "**Y**" here, for simplicity**:** The search algorithm is started by choosing an arbitrary score vector **u** – for example, the column in **Y** with maximum remaining variation. Based on the auxiliary approximation model $\mathbf{X} \approx \mathbf{uw}'$, **X** is regressed columnwise on the *Y*-scores **u** in order to obtain an estimate of the *X*-loading weight vector **w**. Then, based on the approximation model $\mathbf{X} \approx \mathbf{tw}'$, **X** is projected row-wise on the *X*-loading weights **w** in order to calculate the *X*-scores **t**. This can be considered as one local "PCA" NIPALS step in **X**. However, since PLSR is a two-block method, the next NIPALS step is performed in the matrix **Y**: Based on the approximation model, $\mathbf{Y} \approx \mathbf{tq}'$, **Y** is regressed columnwise on the *X*-scores **t** in order to estimate the *Y*-loadings **q**. From the auxiliary model $\mathbf{Y} \approx \mathbf{uq}'$, **Y** is regressed on the *Y*-loadings **q** to update the estimate of the *Y*-scores **u**. As in PCA, this iterative NIPALS procedure of estimating is repeated until convergence, yielding X-loading weight vector **w**, **X**-score vector **t**, **Y**-loadings **q**, and auxiliary **Y**-scores **u**.

For a given PLSR PC, the NIPALS search algorithm in itself is completely symmetric in **X** and **Y**. Accordingly, at convergence the obtained result is completely symmetric: A "sample" score vector **t** and a corresponding "variable" vector **w** are obtained for **X** and a "sample" score vector **u**, and a corresponding "variable" vector **q** are obtained for **Y**. As in PCA, subsequent loading and score vectors are obtained by deflating the matrices **X** and **Y** with respect to the previous component. In PLSR there are two score vectors available for deflation, the score vector **t** of the matrix **X** and the score vector **u** of the matrix **Y**. In order to ensure that **Y** can be predicted from **X**, the

## FRAME 2
## Analysis of Two Data Tables by Partial Least-Squares Regression (PLSR)

The PLSR Data Modeling Process

- Select a set of samples and two sets of variables: *X* variables and *Y* variables.
- Scale each variable to ensure compatible error levels within the *X* variables and within the *Y* variables, and form data tables **X** and **Y** (rows = samples, columns = variables).
- Compute and subtract the mean sample: $\mathbf{X}_{(0)} = \mathbf{X} - \bar{\mathbf{x}}$, $\mathbf{Y}_{(0)} = \mathbf{Y} - \bar{\mathbf{y}}$.
- Model data tables $\mathbf{X}_{(0)}$ and $\mathbf{Y}_{(0)}$ as sum of a few (*A*) trustworthy PLS components as weighted linear combinations of the *X* variables, plus residuals $\mathbf{X}_{(A)}$ and $\mathbf{Y}_{(A)}$. Determine *A*, the number of PCs to trust.
- The parameters may then be collected in matrices of *X*-loading weights $\mathbf{W}_A$, *X*-scores $\mathbf{T}_A$, and *X* and *Y* loadings $\mathbf{P}_A$ and $\mathbf{Q}_A$. The obtained model may be written in many equivalent ways; for example,

$$\begin{aligned}\mathbf{V}_A &= \mathbf{W}_A(\mathbf{P}_A'\mathbf{W}_A)^{-1}\\ \mathbf{T}_A &= (\mathbf{X} - \bar{\mathbf{x}})\mathbf{V}_A\\ \mathbf{X} &= \bar{\mathbf{x}} + \mathbf{T}_A\mathbf{P}_A' + \mathbf{X}_{(A)}\\ \mathbf{Y} &= \bar{\mathbf{y}} + \mathbf{T}_A\,\mathbf{Q}'_A + \mathbf{Y}_{(A)}\end{aligned}$$

  that is,

$$\mathbf{Y} = \mathbf{b}_{0,A} + \mathbf{X}\mathbf{B}_A + \mathbf{Y}_{(A)}$$

  where

$$\mathbf{B}_A = \mathbf{V}_A\mathbf{Q}_A' \quad \text{and} \quad \mathbf{b}_{0,A} = \bar{\mathbf{y}} - \bar{\mathbf{x}}\mathbf{B}_A$$

- Study the obtained parameters and residuals graphically. Look for covariation patterns, clusters; reanalyze without outliers.

The parameters of each PC can be estimated from the residuals after the previous *a* - 1 PCs, $\mathbf{X} = \mathbf{X}_{(a-1)}$ and $\mathbf{Y} = \mathbf{Y}_{(a-1)}$, by the NIPALS algorithm, one PLS component at a time. The PLSR NIPALS algorithm for component #*a* consists of the following steps:

- Choose an arbitrary start score vector **u**.
- $\mathbf{w} = \mathbf{X}'\mathbf{u}/\mathbf{u}'\mathbf{u}$.
- Normalize **w** to length one by $\tilde{\mathbf{w}} = \frac{\mathbf{w}}{\sqrt{\mathbf{w}'\mathbf{w}}}$.
- $\mathbf{t} = \mathbf{X}\tilde{\mathbf{w}}$.
- $\mathbf{q} = \mathbf{Y}'\mathbf{t}/\mathbf{t}'\mathbf{t}$.
- $\mathbf{u} = \mathbf{Y}\mathbf{q}/\mathbf{q}'\mathbf{q}$.
- Iterate until convergence.

Once converged, update **Y** by the deflation:

- $\mathbf{Y} = \mathbf{Y} - \mathbf{t}\mathbf{q}'$.

Project also **X** on **t**, to find the *x*-loading vector **p**, and update **X**:

- $\mathbf{p} = \mathbf{X}'\mathbf{t}/\mathbf{t}'\mathbf{t}$.
- $\mathbf{X} = \mathbf{X} - \mathbf{t}\mathbf{p}'$.

Increment PC counter: $a = a + 1$, and repeat.

deflation is performed asymmetrically in PLSR: The *X*-scores are defined as weighted combinations of the residuals of the *X* variables: $\mathbf{t} = \mathbf{X}_{(a-1)}\mathbf{w}$. These *X*-scores are used for approximating both $\mathbf{X}$ and $\mathbf{Y}$: $\mathbf{Y}_{(a-1)} \approx \mathbf{tq}'$, $\mathbf{X}_{(a-1)} \approx \mathbf{tp}'$. The *Y*-loadings $\mathbf{q}$ were already estimated in the NIPALS process above. But the *X*-loadings $\mathbf{p}$ have to be estimated in a final step, based on the approximation model $\mathbf{X}_{(a-1)} \approx \mathbf{t} \times \mathbf{p}'$, by regression of each variable in $\mathbf{X}_{(a-1)}$ on $\mathbf{t}$.

The deflation is then performed as $\mathbf{X}_{(a)} = \mathbf{X}_{(a-1)} - \mathbf{tp}'$ and $\mathbf{Y}_{(a)} = \mathbf{Y}_{(a-1)} - \mathbf{tq}'$. Then, the next PC is obtained after incrementing the PC counter *a*, by the same NIPALS process. When sufficiently many PCs, $a = 1, 2, \ldots, A$, have been estimated, their parameters are collected in matrices: $\mathbf{W}_A = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_A]$, $\mathbf{T}_A = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_A]$, $\mathbf{Q}_A = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_A]$, $\mathbf{P}_A = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_A,]$. The approximation models with *A* PCs may thus be written $\mathbf{X} = \bar{\mathbf{x}} + \mathbf{T}_A\mathbf{P}_A' + \mathbf{X}_{(A)}$, $\mathbf{Y} = \bar{\mathbf{y}} + \mathbf{T}_A\mathbf{Q}_A' + \mathbf{Y}_{(A)}$. More details about this estimation process are given in Frame 2. It is shown that the *A*-dimensional bilinear model may be summarized by the equivalent linear regression model $\mathbf{Y} = \mathbf{b}_{0,A} + \mathbf{XB}_A + \mathbf{Y}_{(A)}$. Thus, in samples with known values of *X* variables, the *Y* variables can be predicted – not only for the *N* "calibration samples" initially available for estimating the model parameters $\mathbf{b}_{0,A}$ and $\mathbf{B}_A$, but also later on, for new samples of the same general kind.

An important aspect of data-driven modeling like PCA and PLSR is to find the optimal number of PCs to trust, *A*. If too few PCs are used, information is lost, because important structures in the data are left unmodeled and instead treated as residual "noise." This is called "underfitting" or "oversimplification". If too many PCs are used, compared to the reliable information content of the data, the results become statistically unstable and difficult to interpret, because incidental errors are given the same attention as real effects. This is called "overfitting" or "overparameterization." The optimal number of PCs of course depends on the number of patterns, known or unknown – desired or undesired – that vary independently in the system being studied. But, more unexpectedly perhaps, it also depends on the quality and quantity of data available: With precise, high-resolution observations from lots of samples with different patterns of covariation, all the patterns might be picked up by the data-driven modeling. But with imprecise observations from only a few samples that only vary in a few ways, only a few of the true, underlying patterns of co-variation can be picked up by the modeling process. Therefore, like all data-driven modeling, PLSR calls for aggressive experimental planning and humble model interpretation.

In PLSR it is relatively easy to determine the optimal number of PCs, using a resampling procedure such as cross-validation: The modeling process is repeated over and over again, each time keeping part of the sample set as a "secret" test set. The prediction errors $\mathbf{Y}_{(A)} = \mathbf{Y} - (\mathbf{b}_{0,A} + \mathbf{XB}_A)$ in the "secret test samples" decrease until the optimal number of components, $\boldsymbol{A} = \boldsymbol{A}_{\mathrm{Opt}}$, is reached, and then increase again upon overfitting.

The precision of the obtained model parameters, e.g., $\mathbf{B}_{A\mathrm{Opt}}$, may be assessed by the related methods of jack-knifing or boot-strapping: Small elements in $\mathbf{B}_{A\mathrm{Opt}}$ that vary strongly when re-estimated with some samples kept "secret" are deemed unreliable ("not significantly different from zero").[20]

Although PLSR is best known for its ability to establish stable predictive models, which is due to the fact that often only few latent variables are needed for obtaining good predictions, it has been used quite extensively during the recent years in biospectroscopy to study underlying covariance structures in two data sets. It has been used to study covariance structures in mRNA microarray data and FT-IR data of *Campylobacter jejuni* for the investigation of stress responses[21,22] and to study the relation of protein denaturation and water mobility in meat by relating FT-IR microspectroscopic and low field $H^1$ NMR data.[23–25] As in PCA correlation, loading plots can be used to study the correlation of variables with respect to the respective loadings.

*Example.* PLSR was performed for the FT-IR data (720–1200 cm$^{-1}$) as **X**-matrix, with phenotype information as **Y**-matrix: BC, Nisin, serotype, sakacin P values, sakacin P group (indictator variables A,B), and FT-IR group (indicator variables 1,2,3). The preprocessing of the FT-IR data was done as described in the previous section. The phenotype data was scaled by dividing every variable by its standard variation.

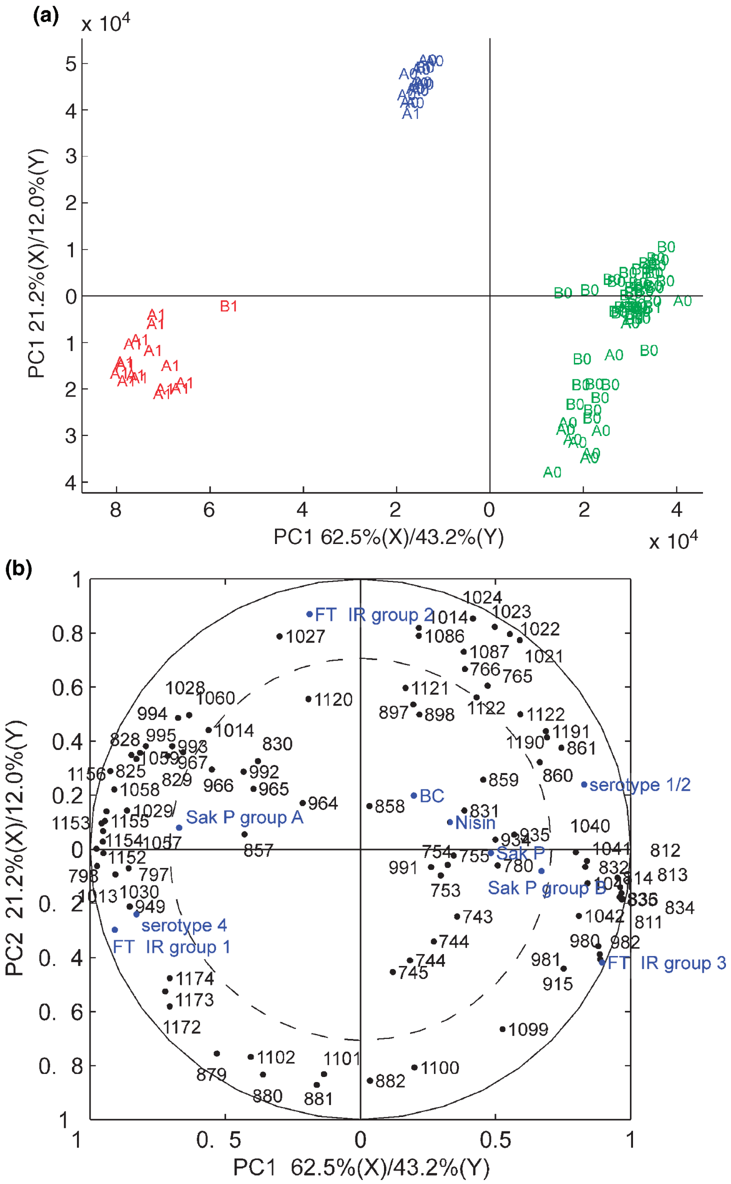

Figure 15.7. Partial least-squares regression with the FT-IR data (720–1200 cm$^{-1}$) as **X**-matrix and phenotype data as **Y**-matrix. (**a**) PLS score plot. (**b**) Correlation loading plot. Color coding is according to Fig. 15.3.

Cross-validation showed that primarily the first two PLS PCs contributed with predictive ability for the $Y$ variables. Hence, these $A = 2$ first PCs are presented here. The $X$-scores $\mathbf{t}$ are shown for the first and the second PLS component in Fig. 15.7a, with symbols and color coding as in Fig. 15.5a. The score pattern in Fig. 15.7a is similar to the PCA score pattern in Fig. 15.5a, but with some variations. In Fig. 15.7b the corresponding PLSR correlation loading plot is shown: The correlations of both the $X$ and $Y$-variables with the first and second PLSR score vector, $\mathbf{t}_1$ and $\mathbf{t}_{2,}$ are plotted along the abscissa and ordinate, respectively. The interpretation of both the score plot and the loading plot is done just as explained for PCA. FT–IR group 1 is seen to be strongly positively correlated to serotype 4 and, for example, 949 $\mathrm{cm}^{-1}$ at about "8 o'clock." FT–IR group 2 at "11 o'clock" is strongly correlated with 1027 $\mathrm{cm}^{-1}$ and strongly anticorrelated with, for example, 881, 882, 883, and 1100, 1101, 1102 $\mathrm{cm}^{-1}$. Sakacin P-group B is somewhat associated with FT–IR group 3 (about 4 o'clock), which is seen to have higher-than-average absorbance at, for example, 980, 981, 982 $\mathrm{cm}^{-1}$ and 934, 935 $\mathrm{cm}^{-1}$. The bands at 835 and 980 $\mathrm{cm}^{-1}$ can be attributed to pyranose rings,[26] and Ref. [15] suggested that the variation in susceptibility toward sakacin P is connected to variations in the cell wall, probably to variations in pyranose.

## 15.4 SIMULTANEOUS ANALYSIS OF SEVERAL DATA BLOCKS BY MULTIBLOCK PCA

As an example for multiblock methods, we would like to discuss multiblock principal component analysis (MPCA), which is a very intuitive extension of PCA toward the analysis of many data sets. MPCA is actually a family of methods, all concerned with simultaneous modeling of several data matrices, but differing with respect to properties of the latent variables. The version discussed in this chapter appears to be especially useful for the treatment of multiblock data including spectroscopy data[27,28]: The obtained block loadings (loadings referring to the $K_\mathrm{b}$ variables in each block matrix $\mathbf{X}^\mathrm{b}$, $b = 1, 2, \ldots, B$) are independent. Thus, the biochemical interpretation of every block loading will be independent from other loadings referring to the same block. As in PCA and in PLSR, MPCA extracts latent variables (block components, defined by block scores $\mathbf{T}_A^b = [\mathbf{t}_1^b, \mathbf{t}_2^b, \ldots, \mathbf{t}_A^b]$ and block loadings $\mathbf{P}_A^b = [\mathbf{p}_1^b, \mathbf{p}_2^b, \ldots, \mathbf{p}_A^b]$) for $A$ PCs in matrix $b = 1,2, \ldots, B$), These may be used for bilinear modeling of each variable in each of matrix: $\mathbf{X}^b \approx \bar{\mathbf{x}} + \mathbf{T}_A^b \mathbf{P}_A^{b'} + \mathbf{X}_{(A)}^b$, where the $\mathbf{X}_{(A)}^b$ are the block's residuals after $A$ PCs. In addition to the block components, also global latent variables $\mathbf{t}_T$ are calculated for each PC to reflect a consensus between all blocks $\mathbf{X}^b$. MPCA has been introduced by Ref. [29], and the corresponding NIPALS algorithm for CPCA is illustrated in Fig. 15.8.

As in PCA and PLSR, loadings and scores are estimated by a version of the NIPALS algorithm. The effect of every estimated component is then subtracted by a deflation step. The NIPALS algorithm is again applied on the deflated matrix, and so on.

The NIPALS algorithm for a given PC is started by choosing an arbitrary score vector, this time an arbitrary estimate of the so-called global score vector $\mathbf{t}_T$ – that is, a score vector of sample properties common to all blocks $b = 1, 2, \ldots, B$. From $\mathbf{t}_T$we obtain block loading vectors $\mathbf{p}^b$ by regressing every block $\mathbf{X}^b$ columnwise on $\mathbf{t}_T$ in order to estimate the block loading vectors $\mathbf{p}^b$. By projecting every block $\mathbf{X}^b$ row-wise onto its own loading, $\mathbf{p}^b$, the set of $B$ block score vectors $\mathbf{t}^b$ are then obtained and collected in the matrix

$$\mathbf{D} = \left[ \mathbf{t}^1 \quad \mathbf{t}^2 \quad \ldots \quad \mathbf{t}^{B'} \right]$$

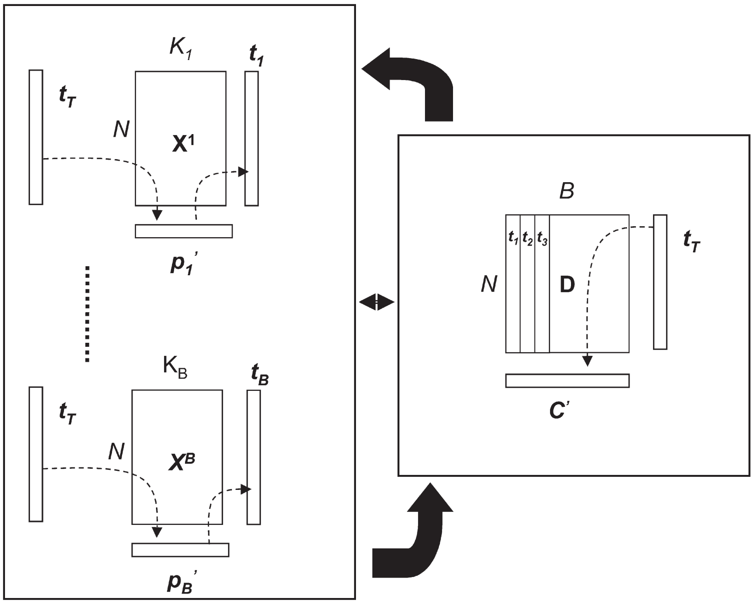

Figure 15.8. Illustration of the NIPALS algorithm (see Frame 3) and the relation between block scores, block loadings, and global scores for MPCA.

Based on the “inner” model $\mathbf{D} \approx \mathbf{t}_T\mathbf{c}_T{}'$, $\mathbf{D}$ is projected onto $\mathbf{t}_T$ in order to obtain the loading weights $\mathbf{c}_T$. Subsequently, $\mathbf{D}'$ is projected onto $\mathbf{c}_T$, and the next estimate for the global score vector $\mathbf{t}_T$ is obtained. Now the block loadings $\mathbf{p}^b$ can be updated, etc. This is repeated until convergence.

In Frame 3 and in Fig. 15.8, it can be seen that the algorithm can be divided into a part where block properties are calculated (Part II) and a part where global properties are considered (Part III).

In MPCA there are several possibilities for deflating the data blocks $\mathbf{X}^b$ – that is, for subtracting components between two NIPALS runs. The most common procedure is to deflate with respect to the variation expressed by the global scores $\mathbf{t}_T$[29,30]:

$$\mathbf{X}_a^b = \mathbf{X}_{a-1}^b - \mathbf{t}_T\mathbf{p}_a^{b'}$$

The advantage of this is that the subtracted variation with respect to the common variation is expressed by the global scores $\mathbf{t}_T$ for this PC. The $A$ global scores are collected in matrix $\mathbf{T}_T = [\mathbf{t}_{1,T}, \mathbf{t}_{2,T}, \ldots, \mathbf{t}_{a,T}, \ldots, \mathbf{t}_{A,T}]$. This deflation procedure also allows reconstructing the global matrix $\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^B]$ by $\mathbf{X} = \mathbf{T}_T\mathbf{P}' + \mathbf{E}$, where $\mathbf{P}$ the matrix of concatenated block loadings is $\mathbf{P}' = [\mathbf{P}_A^{1'}, \mathbf{P}_A^{2'}, \ldots, \mathbf{P}_A^{b'}, \ldots \mathbf{P}_A^{B'}]$, where $\mathbf{p}_k$ are the global loadings obtained by concatenating the block loadings in the same blockorder as $\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^B]$. In this chapter, another possible way for deflation is considered, the deflation on normalized block loadings:

$$\mathbf{X}_{(1)}^b = \mathbf{X}^b - \mathbf{t}_1^b \cdot \tilde{\mathbf{p}}_1^{b'}$$

where $\tilde{\mathbf{p}}_1^b$ is a copy of the block loading, scaled to length 1. This deflation procedure entails a very specific property, namely that the obtained block loadings $\mathbf{p}_b$ are orthogonal. This ensures that there is no correlation between the loadings of different PCs and that different components therefore are independent in interpretative information. After the calculation of

## FRAME 3
## Analysis of Several Single Data Tables by Multiblock Principal Component Analysis (MPCA)

The MPCA Process

- Select a set of samples and several sets of variables for this sample set (rows = samples, columns = variables).
- Scale the variables to comparable noise levels, and form data tables $\mathbf{X}^b$, $b = 1, 2, \ldots, B$.
- Order the sample in every block $\mathbf{X}^b$ (variable set) in the same way (row-to-row correspondence between blocks).
- Compute and subtract the mean sample: $\mathbf{X}^{b,0} = \mathbf{X}^b - \bar{\mathbf{x}}^b$ for every block.
- Scale the blocks $\mathbf{X}^b$ in order to give the blocks comparable weights (e.g., many variables: lower block weight; particularly important variables: higher block weight).
- Model the total data table $\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2 \ldots \mathbf{X}^B]$ as sum of a few trustworthy principal components $a = 1,2, \ldots, A$ plus residual:

$$\mathbf{X} = \bar{\mathbf{x}} + \left[ T_A^1 \tilde{\mathbf{P}}_A^{1'}, \mathbf{T}_A^2 \tilde{\mathbf{P}}_A^{2'}, \ldots, \mathbf{T}_A^b \bar{\mathbf{P}}_A^{b'}, \ldots, \mathbf{T}_A^B \tilde{\mathbf{P}}_A^{B'} \right] + \mathbf{X}_{(A)}$$

where

$$\mathbf{T}_A^b = (\mathbf{X}^b - \bar{\mathbf{x}}^b) \tilde{\mathbf{P}}_A^{b'}$$

Determination of optimal number of PCs, outlier detection, and so on, can be performed as in PCA.

The parameters can be estimated, for example, by the NIPALS algorithm, one component at a time. Given residuals after the previous $a$ - 1 PCs, the NIPALS algorithm for component #a consists of the following steps:

**Part I.** Initialization

- Choose an arbitrary start global score vector $\mathbf{t}_T$.

**Part II.** Computation of block scores and block loadings for each block $b = 1, 2, \ldots, B$:

- $\mathbf{p}^b = \frac{\mathbf{X}^b \mathbf{t}_T}{\mathbf{t}'_T \mathbf{t}_T}$.
- Normalize $\mathbf{p}^\mathrm{b}$ to length 1 by $\tilde{\mathbf{p}}^b = \frac{\mathbf{p}^b}{\sqrt{\mathbf{p}^{b'} \mathbf{p}^b}}$.
- $\mathbf{t}^b = \mathbf{X}^b \tilde{\mathbf{p}}^b$.

**Part III.** Computation of global scores and global loadings

- $\mathbf{D} = \begin{bmatrix} \mathbf{t}^1 & \mathbf{t}^2 & \cdots & \mathbf{t}^B \end{bmatrix}$.
- $\mathbf{c}_T = \frac{\mathbf{D}' \mathbf{t}_T}{\mathbf{t}'_T \mathbf{t}_T}$.
- Normalize $\mathbf{c}_T$ to length 1 by $\tilde{\mathbf{c}}_T = \frac{\mathbf{c}_T}{\sqrt{\mathbf{c}'_T \mathbf{c}}}$.
- $\mathbf{t}_T = \mathbf{D} \tilde{\mathbf{c}}_T$.
- Repeat until convergence.

$A$ components (and thus after $A$ deflation steps), we obtain therefore the linear model of the global matrix $\mathbf{X}$:

$$\mathbf{X} = \bar{\mathbf{x}} + \left[ T_A^1 \tilde{\mathbf{P}}_A^{1'}, \mathbf{T}_A^2 \tilde{\mathbf{P}}_A^{2'}, \ldots, \mathbf{T}_A^b \tilde{\mathbf{P}}_A^{b'}, \ldots, \mathbf{T}_A^B \tilde{\mathbf{P}}_A^{B'} \right]' + \mathbf{X}_{(A)}$$

where $\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^b, \ldots, \mathbf{X}^B]$ is the total data matrix consisting of all data blocks concatenated, the matrices $\mathbf{T}_A^B$ contain $A$ block score vectors, and the matrices $\tilde{\mathbf{P}}_A^b$ contain $A$ normalized block loading vectors for block $b = 1 \ldots B$, while vector $\bar{\mathbf{x}}$ represents the mean of each variable, and $\mathbf{X}_{(A)}$ represents the residuals. The advantage of this model is that it provides more detail on the sample configurations within individual blocks, while all block models still relate to the same latent structures (since the block loadings are the same as the global loadings, except for a trivial rescaling to length 1).

*Example:* For the MPCA illustration, three different ranges of the FT-IR spectra (720–1200 $\mathrm{cm}^{-1}$, fingerprint region and polysaccharide region; 1500–1700 $\mathrm{cm}^{-1}$, protein region; 2800–3000 $\mathrm{cm}^{-1}$, fatty acid regions) were used in order to form three different blocks (preprocessing of FT-IR data was performed as before). In addition, AFLP data and phenotype data were used as two further block matrices. MPCA was performed on the obtained five blocks. The phenotype data block was "passified" (scaled down by a very small number after block standardization) in order to avoid that the phenotype data influence the MPCA model. The advantage of this is that, although the phenotype is not contributing to the model, the phenotype variation can be compared to the remaining four blocks, in its block score space and in the scale-free common correlation loading plot. The AFLP data showed saturation effects in some variable regions. A region without saturation containing 1701 variables was used for the data analysis.

In Fig. 15.9a the score plots (first and second component) for the five different blocks and for the global scores are shown. The symbols, numbers, and color coding are as in Fig. 15.5a. It can be seen that the AFLP data reveal a grouping of the samples/variation pattern similar to that of the FT-IR data in the spectral range 720–1200 $\mathrm{cm}^{-1}$: The AFLP data show three different groups that are clearly separated with respect to serotype and sakacin P susceptibility. In the FT-IR, again three groups are visible (color-coded as FT-IR groups 1, 2, and 3), which are defined by serotype and to some extent by sakacin P susceptibility. The AFLP data and the FT-IR data (720–1200 $\mathrm{cm}^{-1}$) show similar variation patterns, although some of the A and B samples have different FT-IR (720–1200 $\mathrm{cm}^{-1}$) group affiliation. The other spectral regions show some tendencies with respect to sakacin susceptibility, but the separation according to group affiliation (sakacin P group, serotype) is by far not clear. The phenotype block, although down-scaled, shows distinct groups in the score plot, similar to those of FT-IR (720–1200 $\mathrm{cm}^{-1}$). The global scores, combining information from AFLP and the three FT-IR regions (but not the phenotype block), show some grouping, but not nearly as clearly as the AFLP, FT-IR (720–1200 $\mathrm{cm}^{-1}$), and the phenotype blocks.

In Fig. 15.9b the correlation loading plot is shown, revealing relations between three of the data blocks used for MPCA (black: FT-IR range 720–1200 $\mathrm{cm}^{-1}$; green: AFLP data; and blue: reference data) and the global scores $\mathbf{T}$. In general, variables from all data blocks can be shown in the correlation loading plot. For the sake of clarity, only a selection of FT-IR variables is plotted in Fig. 15.9b. It can be seen that some AFLP variables have high correlations with sakacin P susceptibility and FT-IR groups, while others have high correlations with serotype 4. Interpretation of correlation between FT-IR variables and phenotypes are as in the previous sections.

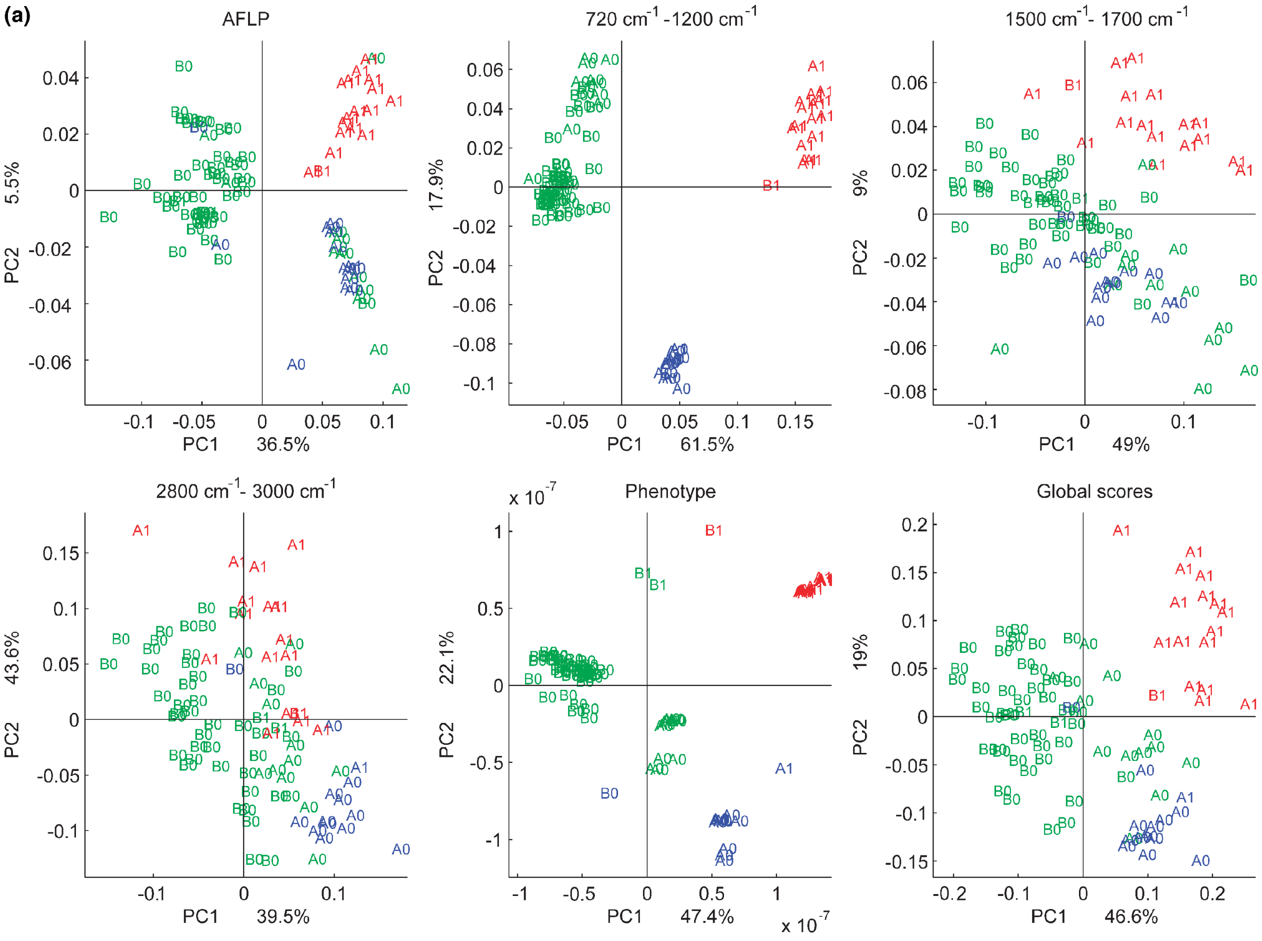

Figure 15.9. Multiblock principal component analysis using AFLP data, different spectral FT-IR regions, and phenotype data as blocks. (**a**) Score plots. (**b**) Correlation loading plot. Color coding is according to Fig. 15.3.

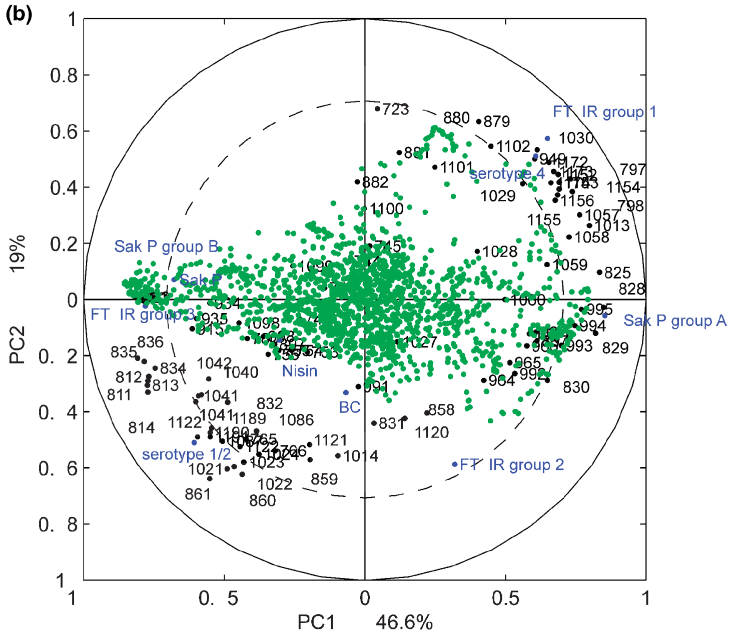

Figure 15.9. (*Continued*)

A general difference between using PCA on every single block and using a multiblock approach is that a multiblock approach is always seeking a "consensus" in the sense that the same variation is appearing in the same components in every block. This makes it possible to compare score results (and other results) directly between the blocks. An important measure for how strong covariation is represented in every block is the explained variance of the block scores. This is very apparent in the block score plot for the FT-IR region from 2800–3000 cm$^{-1}$: The second component has a much higher explained variance (43.6%) than the first component (39.5%). A PCA of this block is changing the order of the axes. The variation with respect to sakacin P and FT-IR group is dominating the "consensus" variation (global scores).

Another comment concerns a difference between MPCA and PLSR as discussed in the previous section: While the multiblock model is influenced equally strong by all not "passified" blocks, in PLSR often the *X*-matrix dominates.

## 15.5 ALTERNATIVE MULTIBLOCK METHODS

Multiblock data analysis has been in focus of intensive research during the last three decades, and many different methods have been discussed in the literature. As a consequence, the reader may be at a loss to decide which method is best suited for a problem. The reader should also be aware of the fact that several methods that are intrinsically similar can be found in the literature under different names or different acronyms. There are also methods of analysis which basically consist of minor variations of standard methods aimed at speeding the computation time or better adapting to a particular domain of interest.

Just like for one- or two-block data analysis, the multiblock methods are usually either of the clustering type, which search for groups of samples that resemble each other, or of the contrasting type (regression or subspace methods – for example, those discussed in this chapter: PCA, PLSR, MPCA), which search for patterns of systematic differences between samples. The multiblock contrasting methods may be divided into three different groups: (1) methods that maximize the correlation between block scores, (2) methods that maximize a common variation pattern, and (3) predictive methods that establish causal models between blocks. In the following we discuss a few examples for each of these three groups:

1. Canonical correlation analysis (CCA) is a method that maximizes the correlation between two blocks of variables and was initially introduced by Hotelling[31,32] into psychometrics. CCA maximizes correlations between block scores and has found many applications in social sciences. Several generalizations of canonical correlation analysis have been proposed for handling situations with more than two blocks of variables; see Refs. 33 and 34 for an overview. Maximizing correlations between blocks scores has certainly disadvantages for biospectroscopic data, where the number of variables is often much higher than the number of samples in every block: Spurious correlations may be found that are not represented very strongly by common underlying variation patterns.
2. MPCA is a method that maximizes a common variation pattern, as we have seen above. Many closely related methods are discussed in literature: In Generalized Procrustes Analysis (GPA),[35] different variation patterns in the different blocks are matched by rotation and scaling. The common structure of the blocks is then investigated by a PCA of the rotated and scaled variation patterns. STATIS[36] is an alternative method where different variation patterns are matched by rotations and scaling as in GPA, but the common variation pattern is calculated in a different way. INDSCAL[37] and Common Components and Specific Weights Analysis (CCSA)[38] weaken the constraint of using one scaling for each block to different scaling factors (weights) for every dimension and block.
3. PLS methods allow the scientist to impose structural relationships between blocks. An example of structural models was already discussed in Section 15.3 (PLSR), where a data matrix **Y** was predicted from another data matrix **X** via PCs **T**. There are many extensions of PLSR – for example, multiblock PLSR, where several blocks are used as $\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^B]$ to predict a data matrix **Y**.[30,39] Another important subgroup of methods for establishing structural equation models are the PLS path models.[40–42] They allow the construction of "predictive networks" between data blocks, where, for example, "feedforward" and "feedback" mechanisms as currently present in systems biology can be built in.

As a conclusion, we can state that most of the methods outlined herein aim at unveiling the hidden patterns and relationships between and within blocks. They also provide important visualizations tools to achieve this purpose. Current research is focused on setting up a general overview of the various methods in order to assess how they relate to each other. Another important issue concerns the validation of the models. As discussed in the context of PLSR, the validation encompasses the assessment of the stability of the model and the uncertainty about the estimated parameters. Validation also consists in setting up a hypothesis testing framework in order to assess the relevance of the patterns revealed by the strategy of analysis.

## REFERENCES

1. H. Martens, M. Martens, 2001. *Multivariate Analysis of Quality: An Introduction*. Chichester, UK: Wiley.
2. A. Kohler, C. Kirschner, A. Oust, H. Martens, 2005. Extended multiplicative signal correction as a tool for separation and characterization of physical and chemical information in Fourier transform infrared microscopy images of cryo-sections of beef loin. *App. Spectros.* **59**: 707–716.
3. S. W. Bruun, A. Kohler, I. Adt, G. D. Sockalingum, M. Manfait, H. Martens, 2006. Correcting attenuated total reflaction-Fourier transform infrared spectra for water vapour and carbon dioxide. *Appl. Spectros.* **60**: 1029–1039.
4. H. Martens, S. W. Bruun, I. Adt, G. D. Sockalingum, A. Kohler, 2007. Correction for temperature – and salt – effects of water in FTIR bio-spectroscopy by EMSC. *J. Chemom.* **20**: 402–417.
5. S. N. Thennadil, H. Martens, A. Kohler, 2006. Physics-based multiplicative scatter correction approaches for improving the performance of calibration models. *Appl. Spectros.* **60**: 315–321
6. H. Wold, 1966. *Nonlinear Estimation by Iterative Least Squares Procedures*, pp. 411–444. London: Wiley.
7. J. M. Farber, P. I. Peterkin, 2000. *Listeria monocytogenes*, In *The microbiological safety of food*, edited by B. M. Lund, T. C. Baird-Parker, G. W. Gould, pp. 1178–1232. Gaithersburg, MD: Aspen Publishers.
8. J. Delves-Broughton, P. Blackburn, R. J. Evans, J. Hugenholtz, 1996. Applications of the bacteriocin, nisin. *Antonie Van Leeuwenhoek* **69**: 193–202.
9. T. Katla, T. Møretrø, I. Sveen, I. M. Aasen, L. Axelsson, L. M. Rørvik, K. Naterstad, 2002. Inhibition of *Listeria monocytogenes* in chicken cold cuts by addition of sakacin P and sakacin P-producing *Lactobacillus sakei*. *J. Appl. Microbiol.* **93**: 191–196.
10. M. Ramnath, M. Beukes, K. Tamura, J. W. Hastings, 2000. Absence of a putative mannose-specific phosphotransferase system enzyme IIAB component in a leucocin A-resistant strain of *Listeria monocytogenes*, as shown by two-dimentional sodium dodecyl sulphate–polyacrylamide gel electrophoresis. *Appl. Environ. Microbiol.* **66**: 3098–3101.
11. M. Ramnath, S. Arous, A. Gravesen, J. Hastings, Y. Héchard, 2004. Expression of *mptC of Listeria monocytogenes* induces sensitivity to class IIa bacteriocins in *Lactococcus lactis*. *Microbiology* **150**: 2663–2668.
12. A. Gravesen, M. Ramnath, K. B. Rechinger, N. Andersen, L. Jänsch, Y. Héchard, J. W. Hastings, S. Knøchel, 2002. High-level resistance to class IIa bacteriocins is associated with one general mechanism in *Listeria monocytogenes*. *Microbiology* **148**: 2361–2369.
13. V. Vadyvaloo, S. Arous, A. Gravesen, Y. Héchard, R. Chaunan-Haubrock, J. Hastings, M. Rautenbach, 2004. Cell-surface alterations in class IIa bacteriocin-resistant *Listeria monocytogenes* strains. *Microbiology* **150**: 3025–3033.
14. T. Katla, K. Naterstad, M. Vancanneyt, J. Swings, L. Axelsson, 2003. Differences in susceptibility of *Listeria monocytogenes* strains to sakacin P, sakacin A, pediocin PA-1, and nisin. *Appl. Environ. Microbiol.* **69**: 4431–4437.
15. A. Oust, T. Moretro, K. Naterstad, G. D. Sockalingum, I. Adt, M. Manfait, A. Kohler, 2006. Fourier transform infrared and Raman spectroscopy for characterization of Listeria monocytogenes strains. *Appl. Environ. Microbiol.* **72**: 228–232.
16. L. M. Rørvik, D. Caugant, M. Yndestad, 1995. Contamination pattern of *Listeria monocytogenes* and other *Listeria* spp. in a salmon slaughterhouse and smoked salmon processing plant. *Int. J. Food Microbiol.* **25**: 19–27.

17. B. Aase, G. Sundheim, S. Langsrud, L. M. Rørvik, 2000. Occurrence of and a possible mechanism for resistance to a quarternary ammonium compound in *Listeria monocytogenes*. *Int. J. Food Microbiol.* **62**: 57–63.
18. H. Martens, J.P. Nielsen, S. B. Engelsen, 2003. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. *Anal. Chem.* **75**: 394–404.
19. H. Martens, T. Næs, 1989. *Multivariate Calibration*. Chichester: Wiley & Sons.
20. F. Westad, H. Martens, 2000. Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression *J. Near Infrared Spectrosc.* **8**: 117–124.
21. A. Oust, B. Moen, H. Martens, K. Rudi, T. Naes, C. Kirschner, A. Kohler, 2006. Analysis of covariance patterns in gene expression data and FT-IR spectra. *J. Microbiol. Methods* **65**: 573–584.
22. B. Moen, A. Oust, O. Langsrud, N. Dorrell, G.L. Marsden, J. Hinds, A. Kohler, B. W. Wren, K. Rudi, 2005. Explorative multifactor approach for investigating global survival mechanisms of Campylobacter jejuni under environmental conditions. *Appl. Environ. Microbiol.* **71**: 2086–2094.
23. H. C. Bertram, A. Kohler, U. Böcker, R. Ofstad, H. J. Andersen, 2006. Heat-induced changes in myofibrillar protein structures and myowater of two pork qualities. A combined FT–IR spectroscopy and low-field NMR relaxometry study. *J. Agric. Food Chem.* **54**: 1740–1746.
24. U. Böcker, R. Ofstad, H. C. Bertram, B. Egelandsdal, A. Kohler, 2006. Salt-induced changes in pork myofibrillar tissue investigated by FT–IR microspectroscopy and light microscopy. *J. Agric. Food Chem.* **54**: 6733–6740.
25. Z. Y. Wu, H. C. Bertram, A. Kohler, U. Böcker, R. Ofstad, H. J. Andersen, 2006. Influence of aging and salting on protein secondary structures and water distribution in uncooked and cooked pork. A combined FT–IR microspectroscopy and H-1 NMR relaxometry study. *J. Agric. Food Chem.* **54**: 8589–8597.
26. G. Socrates, 2001. *Infrared and Raman Characteristic Group Frequencies: Tables and Charts*. Chichester: Wiley.
27. D. Chessel, M. Hanafi, 1996. Analyses de la co-inertie de *K* nuages de points. *Rev. Stat. Appl.* **XLIV**: 35–60.
28. M. Hanafi, G. Mazerolles, E. Dufour, E. M. Qannari, 2006. Common components and specific weight analysis and multiple coinertia analysis applied to the coupling of several measurement techniques. *J. Chemom.* **20**: 172–183.
29. S. Wold, S. Hellberg, Y. Lundstedt, M. Sjostrom, H. Wold, 1987. Proceedings, Symposium on *PLS Model Building*: Theory and Application, Frankfurt am Main.
30. J.A. Westerhuis, T. Kourti, J. F. Macgregor, 1998. Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemometrics* **12**: 301–321.
31. H. Hotelling, 1935. The most predictable criterion *J. Educ. Psychol.* **26**: 139–142.
32. H. Hotelling, 1936. Relations between two blocks of variates. *Biometrika* **28**: 321–377.
33. M. Hanafi, H. A. L. Kiers, 2006. Analysis of K blocks of Data with differentiel emphasis on agreement between and within blocks. *Comput. Stat. Data Anal.* **51**: 1491–1508.
34. J. R. Kettenring, 1971. Canonical analysis of several blocks of variables. *Biometrika* **58**: 433–451.
35. J. C. Gower, 1975. Generalised Procrustes analysis. *Psychometrika* **40**: 33–51.
36. C. Lavit 1988. *Analyse conjointe de tableaux quantitatifs*. Paris: Masson.
37. J. D. Carroll, J. J. Chang, 1970. Analysis of individual differences in multidimensional scaling via an *N*-way generalisation of "Eckart–Young" decomposition. *Psychometrika* **35**: 283–319.
38. E. M. Qannari, I. Wakeling, P. Courcoux, J. H. MacFie, 2000. Defining the underlying sensory dimensions. *Food Qual. Preference* **11**: 151–154.

39. J. A. Westerhuis, A. K. Smilde, 2001. Deflation in multiblock PLS. *J. Chemom.* **15**: 485–493.
40. J. B. Lohmöller, 1989. *Latent Variables Path Modeling with Partial Least Squares*. Heidelberg: Physica-Verlag.
41. M. Tenenhaus, J. P. Gauchi, C. Ménardo, 1995. Régression PLS et applications (PLS regression and applications). *Rev. Stat. Appl.* **43**: 7–63.
42. H. Wold, 1982. Soft modelling: the basic design and some extensions. In *System under Indirect Observation*, edited by K. G. Jöreskog, H. Wold, pp. 1–54. Amsterdam: North Holland.