The Handbook of Metabonomics and Metabolomics
John C. Lindon, Jeremy K. Nicholson and Elaine Holmes (Editors)
© 2007 Published by Elsevier B.V.

Chapter 6

Chemometrics Techniques for Metabonomics

Johan Trygg[1] and Torbjörn Lundstedt[2,3]

[1]*Research group for Chemometrics, Institute of Chemistry, Umeå University, Sweden*
[2]*Department of Pharmaceutical Chemistry, Uppsala University, Sweden*
[3]*AcurePharma, Uppsala, Sweden*

6.1. Introduction

In biology, as well as in other branches of science and technology, there is a steady trend towards the use of more variables (properties) to characterize observations (e.g. samples, experiments, time points). Often, these measurements can be arranged into a data table, where each row constitutes an observation and the columns represent the variables or factors we have measured (e.g. intensities at a specific wavelength, mass-to-charge ratio, NMR chemical shift). This development generates increasingly complex data tables, which are hard to summarize and overview without appropriate tools. Thus, in this chapter we will try to guide the reader through a chemometrical approach for extracting information out of data.

Chemometrics is an established field in data analysis [1–3] and has proven valuable in the analysis of "omics" data in many applications [4–10]. It includes efficient and robust methods for modelling and analysis of complicated chemical/biological data tables that produce interpretable and reliable models capable to handle incomplete, noisy and collinear data structures. These methods include principal component analysis [11] (PCA) and partial least squares [12–15] (PLS). Chemometrics also provides a means of collecting relevant information through statistical experimental design [16–18]. Therefore, chemometrics can be defined as the information aspect of complex biological and chemical systems.

Chemometrics has grown into a well-established data analysis tool in areas such as multivariate calibration [19, 20] quantitative structure-activity modeling [21, 22], pattern recognition [23–25] and multivariate statistical process monitoring and control [26–28]. Although seemingly diverse disciplines, the common denominator in these application areas are that high complexity data tables are generated and that these can be analysed and interpreted by means of chemometric methods. However, in biology, chemometric methodology has been largely overlooked in favour of traditional statistics. It is not until recently that the overwhelming size and complexity of the "omics" technologies has driven biology towards the adoption of chemometric methods.

There are two main categories of metabonomic studies:

1. Class specific studies, for example disease diagnosis or toxicological classification
2. Dynamic studies, for example the temporal progression of a treatment.

The common theme is that design of experiments (DOE) is used in combination with multivariate analysis (MVA). A brief introduction to the chemometrical approach, DOE and MVA will be given and later illustrated by an example.

6.1.1. Making data contain information – Design of Experiments

The metabonomics approach is more demanding on the quality, accuracy and richness of information in data sets. The DOE [16, 17] is recommended to be used through the whole process, from defining the aim of the study to the final extraction of information.

The objective of experimental design is to plan and conduct experiments in order to extract the maximum amount of information in the fewest number of experimental runs. The basic idea is to devise a small set of experiments, in which all pertinent factors are varied systematically. This set usually does not include more than 10 to 20 experiments. By adding additional experiments, one can investigate factors more thoroughly, for example the time dependence from 2 to 5 time points. In addition, the noise level is decreased by means of averaging, the functional space is efficiently mapped and interactions and synergisms are seen.

6.1.2. Extracting information from data – Overview and classification

In metabonomic studies, the observations and samples are often characterized using modern instrumentation such as gas chromatography-mass spectrometry (GC-MS), liquid chromatography-mass spectrometry (LC-MS) and nuclear magnetic resonance (NMR) spectroscopy. The analytical platform is important and largely determined by the biological system and the scientific question. Multivariate analyses based on projection methods represent a number of efficient and useful methods for the analysis and modelling of these complex data. The PCA [11] is the workhorse in chemometrics. Using PCA it is possible to extract and display the systematic variation in the data. A PCA model provides a summary or overview of all observations or samples in the data table. In addition, groupings, trends and outliers can also be found. Hence, such projection-based methods represent a solid basis for metabonomic analysis. Canonical correlation [29], correspondence analysis [30], neural networks [31, 32], Bayesian modeling [33] and hidden Markov models [34] represent additional modelling methods but are outside the scope of this chapter.

6.1.3. Investigating complicated relationships – Discrimination and prediction

Metabonomic studies typically constitute a set of controls and treated samples, including additional knowledge of the samples, for example dose, age, gender and diet. In these situations, it is possible for a more focussed evaluation and analysis of the data. That is, rather than asking the question "what is there?", one can start to ask, "what is its relation to?" or "what is the difference between?". In modelling, this additional knowledge constitutes an extra data table, that is a Y matrix. The (PLS) [14] and Orthogonal-PLS [35–38] (OPLS) represent two modelling methods for relating two data tables. The Y data table can be both quantitative (e.g. age, dose concentration) and qualitative (e.g. control/treated) data.

6.2. Chemometric approaches to metabonomic studies

The underlying philosophy of chemometrics in combination with the chemometric toolbox can efficiently be applied throughout a metabonomic study. This philosophy is useful from the start of a study (defining the aim) through the whole process to the biological interpretation. This strategy is described step by step below.

6.2.1. Step 1 Definition of aim

It is important to formulate the objectives and goals of the metabonomic study.

- What is previously known?
- What additional information is needed to be known?
- How to reach the objectives, that is what experiments are needed and how to perform them?

6.2.2. Step 2 Study design

6.2.2.1. Class specific studies

The traditional approach to metabonomic disease diagnosis is to identify a group of control observations and another group of observations known to have a specific disease. What is not taken into account is that they may have other, not diagnosed diseases or conditions. Hence, in modelling, disease diagnosis can be regarded as either a two-class or a one-class problem.

Two-class problem: Disease and control observations define two separate classes.
One-class problem: Only disease observations define a class, control samples are too heterogeneous, for example due to other variations caused by diseases, gender, age, diet, lifestyle, genes, unknown factors and so on.

6.2.2.2. Dynamic studies

Metabonomic studies that involve the quantification of the dynamic metabolic response are best evaluated using sequential sampling over an appropriate time course. The evaluation of human biofluid samples is further complicated by a high degree of normal physiological variation caused by genetic and lifestyle differences. Dynamic sampling makes it possible to evaluate and handle the different types of variations such as individual differences in metabolic kinetics, circadian rhythm and fast and slow responders.

6.2.3. Step 2a – Selection of objects

The selection of the objects (e.g. individuals, rats or plants) needs to span the experimental domain in a balanced and systematic manner. To be able to do this, we have to characterize the objects with both measured and observed descriptors. This often includes setting up specific inclusion and exclusion criteria for the study, such as age span (e.g. 18–45 years), body mass index (e.g. 20–30), medicinal chemistry profiles (e.g. lipids, glucose), gender, tobacco habits and use of drugs. In addition to those criteria, additional information regarding each object is collected by questionnaires that include life style factors, food and drinking habits, social situation and so on. This collected information represents a multivariate profile (with K descriptors) for each object that is a fingerprint of its inherent properties.

Geometrically, the multivariate profile represents one point in K-dimensional space, whose position (coordinates) in this space is given by the values in each descriptor. For multiple profiles, it is possible to construct a two-dimensional data table, an X matrix, by stacking each multivariate profile on top of each other. The N rows then produce a swarm of points in K-dimensional space, see Figure 6.1.

6.2.3.1. Projection-based methods

The main, underlying assumption of projection-based methods is that the system or process under consideration is driven by a small number of latent variables (LVs) [39]. Thus projection-based methods can be regarded as a data analysis toolbox, for indirect observation of these LVs. This class of models are conceptually very different from traditional regression models with independent predictor variables.

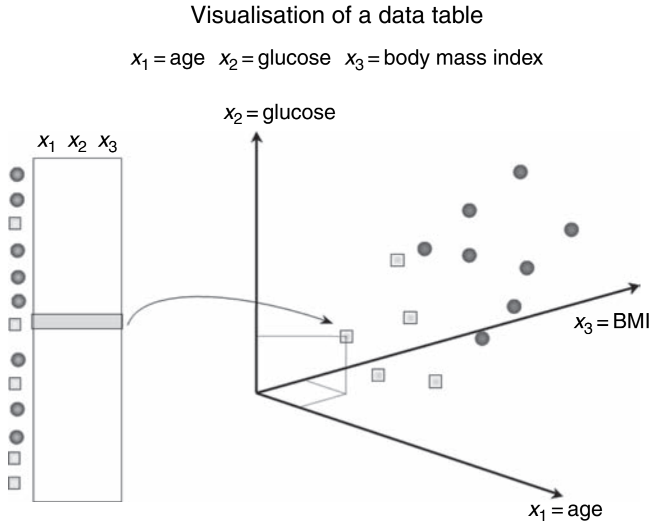

Visualisation of a data table
x_1 = age x_2 = glucose x_3 = body mass index
x_2 = glucose
x_1 x_2 x_3
x_3 = BMI
x_1 = age

Figure 6.1. Each row (e.g. object or observation) in a K-dimensional data table (here with $K = 3$ variables, designated x_1, x_2, x_3) can be represented as a point in a K-dimensional space (here one point in a three-dimensional space). The coordinates for each object in this multi-dimensional space are given by its three variables, that is a multivariate profile. A data table with N rows then corresponds to a swarm of points. Points that are close to each other have more similar properties than points that lie far apart.

They are able to handle many, incomplete and correlated predictor variables in a simple and straightforward way, hence their wide use.

Projection methods convert the multi-dimensional data table into a low-dimensional model plane that approximates all rows (e.g. objects or observations) in X, that is the swarm of points. The first PCA model component $(t_1 p_1{}^T)$ describes the largest variation in the swarm of points. The second component models the second largest variation and so on. All PCA components are mutually linearly orthogonal, see Figure 6.2. The scores (T) represent a low-dimensional plane that closely approximates X, that is the swarm of points. A scatter plot of the first two score vectors $(t_1 - t_2)$ provides a summary or overview of all observations or samples in the data table. Groupings, trends and outliers are revealed. The position of each object in the model plane is used to relate objects to each other. Hence, objects that are close to each other have a similar multivariate profile, given the K descriptors. Conversely, objects that lie far from each other have dissimilar properties.

Analogous to the scores, the loading vectors (p_1, p_2) define the relation among the measured variables, that is the columns in the X matrix. A scatter plot, also known as the *loading plot*, shows the influence (weight) of the individual X-variables in the model. An important feature is that directions in the score plot correspond to

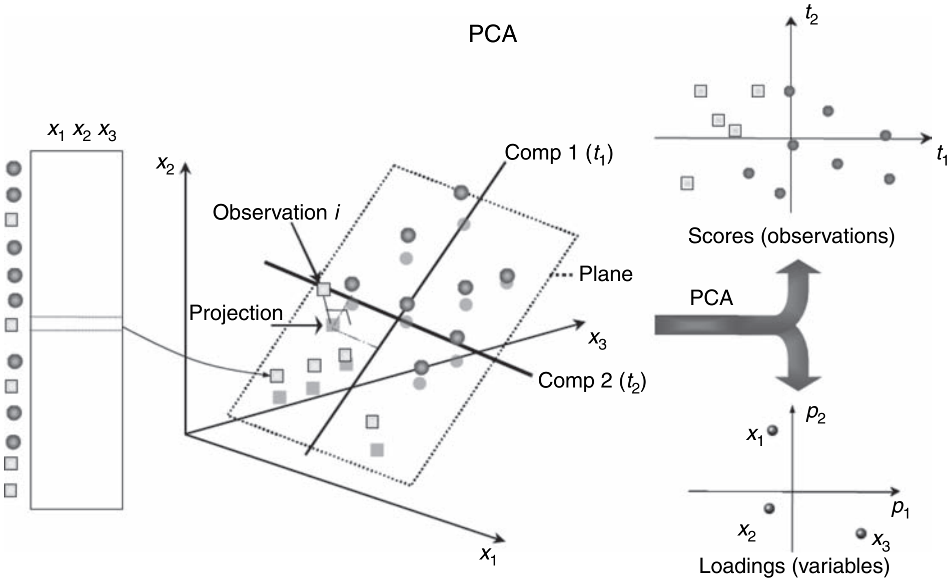


Figure 6.2. A principal component analysis (PCA) model approximates the variation in a data table by a low dimensional model plane. This model plane represents a two-dimensional projection of the multi-dimensional data and provides a score plot, where the relation among the observations or samples in the data table is visualized, for example if there are any groupings, trends or outliers. The loadings plot describes the influence of the variables and the relation among them. An important feature is that directions in the score plot correspond to directions in the loading plot, and vice versa.

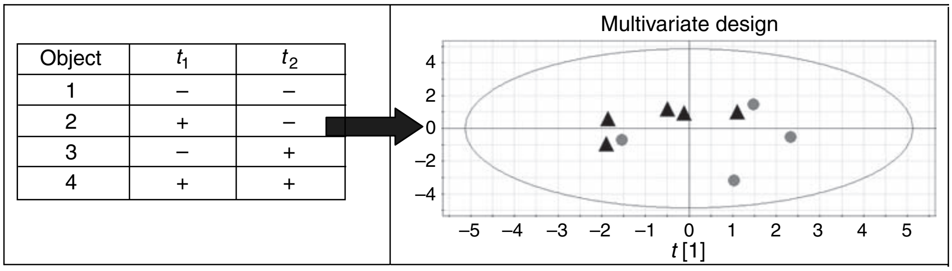
directions in the loading plot, for example for identifying which variables (loadings) separate different groups of objects (the scores). This is a powerful tool for understanding the underlying patterns in the data. Hence, projection-based methods represent a solid basis for metabonomic analysis.

The part of X that is not explained by the model forms the residuals (E) and represents the distance between each point in K-space and its projection on the plane. The scores, loadings and residuals together describe all of the variation in X.

$$X = \mathrm{TP}^T + E = t_1 {p_1}^T + t_2 {p_2}^T + E$$

6.2.3.2. Multivariate design

The need and usefulness of experimental design in complex systems should be emphasized, because it creates a controlled setting of the environment even though most of the variation between the different objects is uncontrolled. Multivariate design (MVD) [40, 41] is a combination of multivariate characterisation (MVC) [42–44], principal component analysis (PCA) and Design of Experiments (DOE) to select a diverse set of objects that represent all objects, that is spans the variation.



Object	t_1	t_2
1	–	–
2	+	–
3	–	+
4	+	+

Figure 6.3. Four objects are selected according to a multivariate design that span the model variation.

There is a number of different experimental designs that can be applied to span the variation in a systematic way and to obtain well-balanced data. The most commonly used are factorial designs [17] and D-optimal design [45] that fulfil the criteria of balanced data and orthogonality. In MVD, the principal component model scores, for example, t_1 and t_2 are used to select the objects, see Figure 6.3. The selection is based on diversity between the objects.

6.2.4. Step 2b Dynamic sampling

Biological processes are dynamic by nature, that is there is a temporal progression. Some problems are caused by quick and slow responders following intervention or treatment. For this reason, the study design is laid out as sequential samples over an appropriate time course to capture individual trajectories. Sampling period and interval is based on the expected or known pharmaco-kinetics of the expected effect. In other words, design of experiments is used to maximize the information content and increase the chances of capturing all possible variations of responses. This allows flexibility to the subsequent analysis and an unbiased evaluation of each individual's kinetic profile. This also implies that the often assumed control (or pre-dose) and treated modelling approach is not optimal, as it fails to take into account the individual dynamics, for example slow and fast responders. In addition, for dynamic studies the traditional control group does not exist. Instead, each individual (object) is its own reference control.

6.2.5. Step 3 Sample preparation and characterisation

In metabonomics, it is important to keep the experimental and biological variation at a minimum. At the same time, the metabolic analysis should be global, quantitative, robust, reproducible, accurate and interpretable. In addition, the physico-chemical diversity of metabolites (amino acids, fatty acids, carbohydrates and organic acids) raises problems for extraction and working up procedures for different analytical techniques. Here, design of experiments represents an important strategy to systematically investigate factors and optimize the experimental protocols. Typical working up procedures for NMR spectroscopy for biofluids and tissue extraction is found in Appendix 4, in the SMRS Policy document [46]. For GC-MS, see References [4, 5].

6.2.6. Step 4 Evaluation of the collected data

In contrast to a ^{1}H-NMR spectrum, data collected from hyphenated instruments such as GC-MS, LC-MS and UPLC-NMR must be processed more extensively before multivariate analysis. The reason is the two-dimensional nature (e.g. chromatogram/mass spectra) of the data for each sample. Curve resolution or deconvolution methods are mainly applied for data processing [47–50] that result in a multivariate profile for each sample. Since a variable in a data table should define the same property over all samples, variability in NMR peak shifts also cause problems for statistical modelling. Because of this, a multitude of different peak alignment methods have been developed [51, 52]. Typically, alignment methods rely upon having a master or reference profile.

Projection-based methods are sensitive to scaling of the variables. Scaling of variables changes the length of each axis in the K-dimensional space. The primary objective of scaling is to reduce the noise in the data, and thereby enhance the information content and quality. Column centring, whereby the mean trajectory is removed from the data, is followed by either no scaling or pareto scaling of the variables. Pareto scaling is recommended for metabonomic data and is done by dividing each variable by the square root of its standard deviation.

Principal component analysis is used to get an overview of the multivariate profiles. Examining the scatter plot of the first two score vectors $(t_1 - t_2)$ reveals the homogeneity of the data, any groupings, outliers and trends. Strong outliers are found as deviating points in the scatter plot. The Hotelling's T2 region, shown as an ellipse in Figure 6.4 (left), defines the 95% confidence interval of the modelled variation [53]. Outliers may also be detected in the model residuals. The distance to model plot [3] (DModX) can be used and is a statistical test for detecting outliers based on the model residual variance, see Figure 6.4 (right).

Interesting individual observations such as outliers can be examined and interpreted by the contribution plot [54]. It displays the weighted difference between the observation and the model centre. Hence, we can identify what is unique (deviating) for an observation compared to "normality". Similarly, the contribution plot can also be used for comparing different observations.

In the scores plot, Figure 6.5, two groupings are observed (yellow boxes and blue circles). Examining the scatter plot of the first two loading vectors $(p_1 - p_2)$ reveals the relation among the variables. In addition, directions in the scores plot correspond

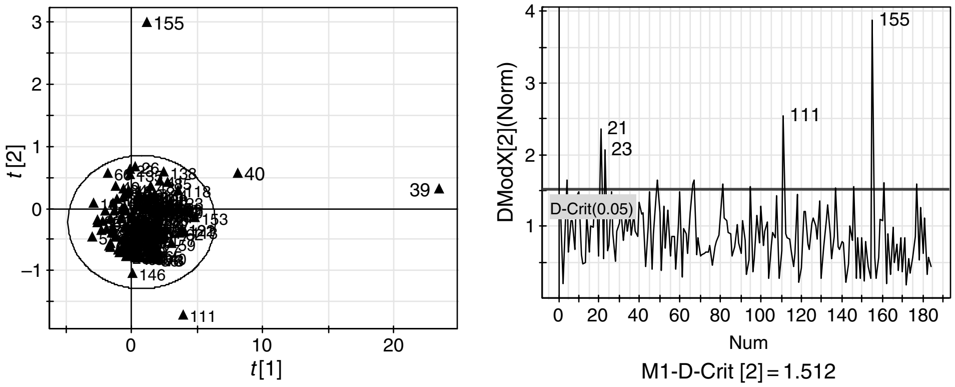


Figure 6.4. In the score plot (left figure), the model is defined by the Hotelling's T^2 ellipse (95% confidence interval) and observations outside the confidence ellipse are considered outliers. Outliers can also be detected by the distance to model parameter, DModX, based on the model residuals (right figure).

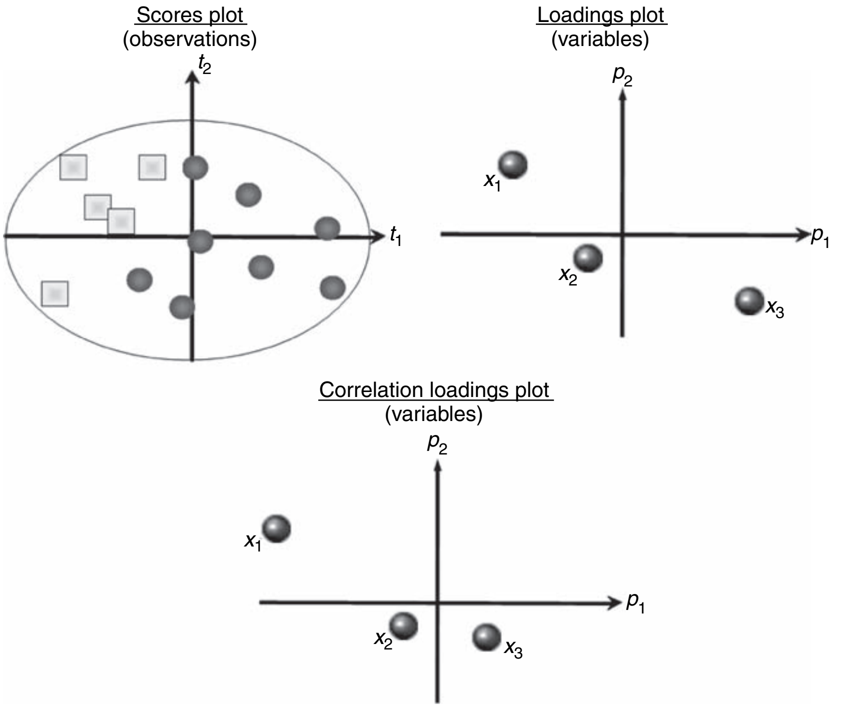


Figure 6.5. The scores plot, the loadings plot and the correlation loadings plot are shown for the first two model components. The scores plot displays an overview of the relationship between the observations (e.g. samples). The loadings plot shows the covariance between each individual variable and the score components. The correlation loadings plot display the correlation between each variable and the score components.

to directions in the loadings plot. This provides the ability to interpret the way the variables are related to a pattern or observations found in the scores plot. This is a powerful tool for understanding the underlying structures in the data.

In Figure 6.5. the loadings plot shows that the variable x_1 is positively correlated to the group marked with yellow boxes, and negatively correlated to the group marked with blue circles. Conversely, variable x_3 is positively correlated to the group with blue circles and negatively correlated to the group marked with yellow boxes. A complementary plot to the loadings plot is the correlation loadings plot in Figure 6.5. It reveals the correlation of each variable to the score components in the model. The correlation loadings plot is not dependant upon the scale or size of the variable contrary to the loadings plot.

The loadings plot in Figure 6.5 shows that the x_1 variable has a similar distance from the origin as the x_3 variable. However, in the correlation loadings plot, the x_1 variable has a stronger correlation than the x_3 variable. This means that the x_1 variable has both strong covariance (given by the loadings plot) and strong correlation (given by the correlation loadings plot) with the first two model score components, compared to the x_3 variable.

Compared to the loadings plot, the correlation loadings plot is scale independent.

The prior knowledge gained in Step 2 (Study Design) gives us the ability to separate the observations in at least two different classes. For instance, observations diagnosed with disease vs another group of observations not having the disease. However, knowledge of different types of variations in the collected data can be handled either separately or jointly.

6.2.7. Soft Independent Modelling of Class Analogy

The Soft Independent Modelling of Class Analogy (SIMCA) [25] method is a supervised classification method based on PCA. The idea is to construct a separate PCA model for each known class of observations. These PCA models are then used to assign the class belonging to observations of unknown class origin by the prediction of these observations into each PCA class model where the boundaries have been defined by the 95% confidence interval. Observations that are poorly predicted by the PCA class model, hence have large residuals, are classified being outside the PCA model and do not belong to the class.

The SIMCA model, as shown in Figure 6.6 (left), illustrates only one class of observations with strong homogeneity and is well modelled by PCA. This is commonly referred to as the asymmetric case. In Figure 6.6 (right), there are two homogenous classes of observations, each separately modelled by PCA. New observations are predicted into each model, and assigned as belonging to either of the classes, none of the classes or both of the classes.

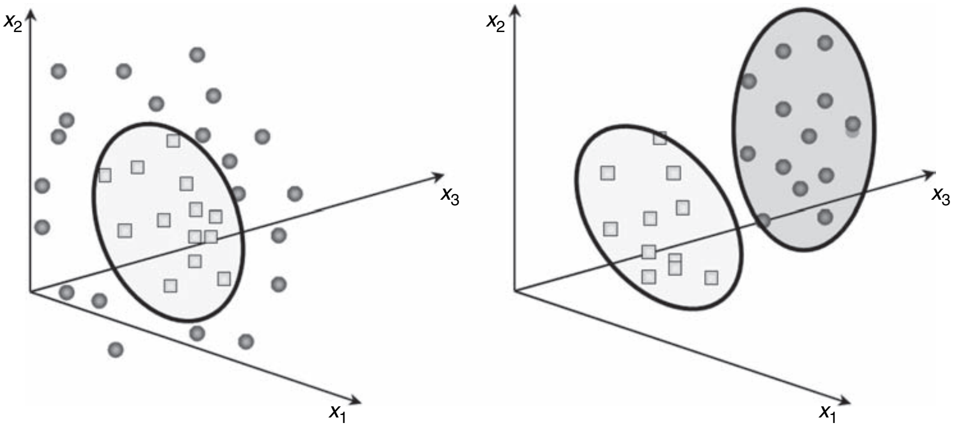


Figure 6.6. Illustration of SIMCA classification. In the left figure, the one class classifier is shown, referred to as the asymmetric case. In the right figure, the SIMCA classification is shown with two classes, separately modelled by PCA.

6.2.8. Partial least squares method by projections to latent structures

~~The~~ PLS [12–15] is a method commonly used where a quantitative relationship between two data tables X and Y is sought between a matrix, X, usually comprising spectral or chromatographic data of a set of calibration samples, and another matrix, Y, containing quantitative values, for example concentrations of endogenous metabolites (Figure 6.7). The PLS can also be used in discriminant analysis, that is PLS-DA. The Y matrix then contains qualitative values, for example class belonging, gender and treatment of the samples. The PLS model can be expressed by:

$$\text{Model of } X: \qquad X = \mathrm{TP}^T + E$$

$$\text{Model of } Y: \qquad Y = \mathrm{TC}^T + F$$

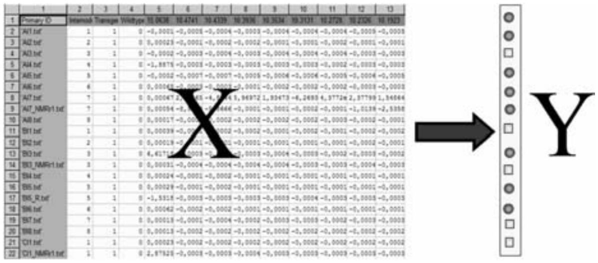


Figure 6.7. Class information can also be used to construct an additional matrix, hereinafter called the Y matrix, consisting of a discrete ‘dummy’ variable where [1]/[0] indicates the class membership.

The PLS models are negatively affected by systematic variation in the X matrix that is not related to the Y matrix, that is that is not part of the joint correlation structure between $X-Y$. This leads to some pitfalls regarding interpretation and has potentially major implications in our selection of metabolite biomarkers, for example positive correlation patterns can be interpreted as negligible or even become negative.

6.2.9. The Orthogonal-PLS method

The OPLS [35] method is a recent modification of the PLS method [14]. The main idea of OPLS is to separate the systematic variation in X into two parts, one that is linearly related to Y and one that is unrelated (orthogonal) to Y. This partitioning of the X-data facilitates model interpretation and model execution on new samples [35]. The OPLS model comprises of two modelled variations, the Y-predictive $(T_\mathrm{p}P_\mathrm{p}^T)$ and the Y-orthogonal $(T_\mathrm{o}P_\mathrm{o}^T)$ components. Only the Y-predictive variation is used for the modelling of Y $(T_\mathrm{p}C_\mathrm{p}^T)$.

$$\text{Model of } X: \qquad X = T_\mathrm{p}P_\mathrm{p}^T + T_\mathrm{o}P_\mathrm{o}^T + E$$

$$\text{Model of } Y: \qquad Y = T_\mathrm{p}C_\mathrm{p}^T + F$$

E and F are the residual matrices of X and Y respectively. The OPLS can, analogously to PLS-DA, be used for discrimination (OPLS-DA), see, for instance, Reference [55]. In Figure 6.8, it is shown how additional knowledge, the Y matrix (e.g. gender), is used in the modelling to identify directions in the X model that relate X to Y.

Example study: A food supplement study with dynamic sampling

Step 1 Definition of aim

Investigate the effects on humans of a food supplement by NMR-based metabonomics. [This needs a few sentences to expand the study details, for example nature of the diet, how many patients, whether plasma or serum was used, type of NMR spectra – CPMG or other type?]

Step 2 Study design

Potential study objects were screened two weeks before the start of the study with a number of inclusion and exclusion criteria (e.g. gender, BMI, age, clinical chemistry) and a questionnaire to provide more in-depth information about lifestyle habits. The information was collected as a multivariate profile of each individual. A PCA was performed on the collected data, followed by an experimental design to select a

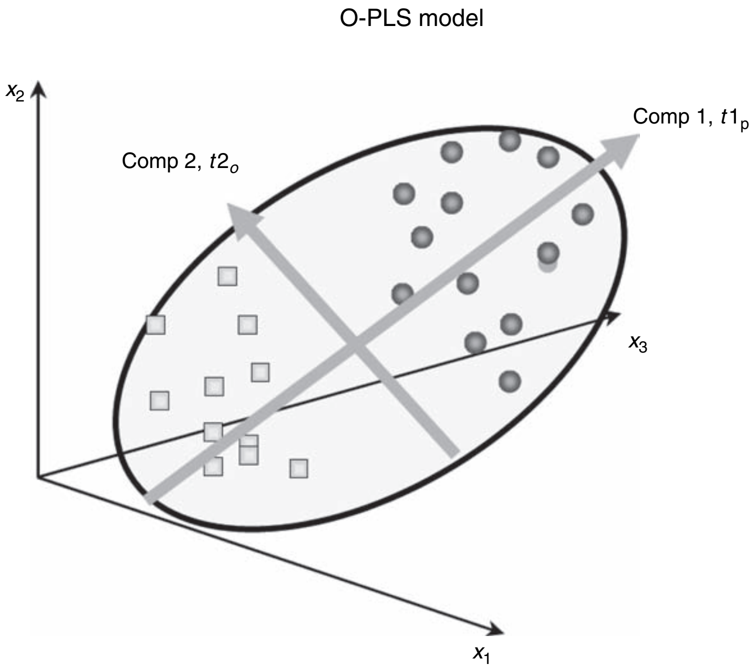


Figure 6.8. A geometrical illustration of the OPLS-DA model. Component 1 ($t1_p$) is the predictive component and displays the between-class ([blue circles], [yellow squares]) variation of the samples. The corresponding loading profile can be used for identifying variables important for the class separation. Component 2 ($t2_o$) is the Y-orthogonal component and models the within group (intra-class) variation.

diverse set of objects. Four objects were selected, in agreement with a multivariate design; see Figure 6.3, for a deeper analysis of a few specific endogenous metabolites. A dynamic study design was laid out for all of the objects whereby a blood sample was withdrawn at each visit and the plasma prepared, see Figure 6.9.

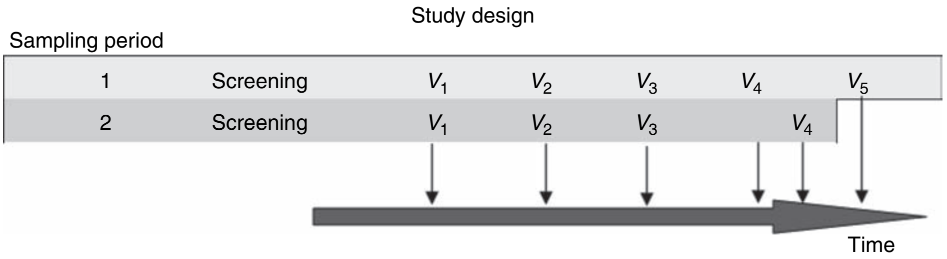

Study design
Sampling period
1 Screening V_1 V_2 V_3 V_4 V_5
2 Screening V_1 V_2 V_3 V_4
Time

Figure 6.9. Four or five sampling times were set up for the two different sampling periods. The dynamic sampling increases the detection of effect and object differences in metabo-kinetics, for example from slow and fast responders.

Step 3 Sample preparation and characterization

Working up and sample preparation was done according to Standard Operating Procedures (SOP), see References [4, 5, 46].

Step 4 Evaluation of the collected data

Prior to all modelling, column centering was applied to the NMR spectral data. Following this, a PCA model was calculated, to obtain an overview of the data. The scores plot (t_1-t_2), shows a summary of all samples and this clearly separates the two different sampling periods, see Figure 6.10.

The corresponding loading plot (p_1-p_2) indicated that there was a problem with peak alignment between the two sampling periods. A line plot of all NMR spectra confirms that this was indeed the case. In addition to the alignment problem, there are also major amplitude differences, see Figure 6.11. Alignment methods can be used to correct for the differential chemical shifts observed. Here, a covariance alignment method was applied.

Following alignment, a new PCA model showed that there still was a separation between each of the two sampling periods, although with minor overlap. Subtracting the screening NMR spectrum from each individual can reduce the amplitude differences between the sampling periods. This is due to the fact that we are interested

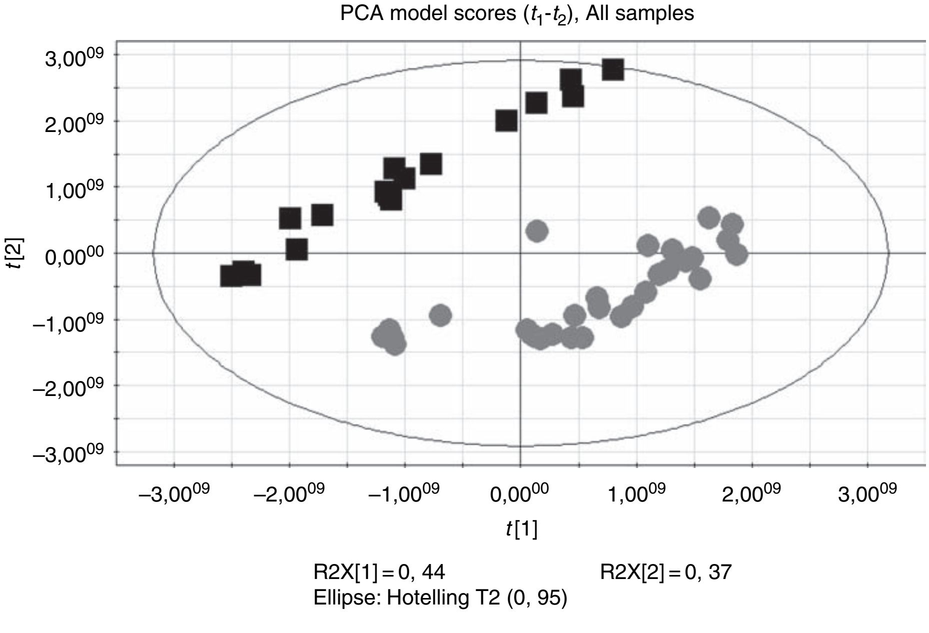


Figure 6.10. PCA scores plot (t_1-t_2) of the NMR spectra shows a clear separation between the sampling periods where black squares represent the first sampling period, and red circles the second period.

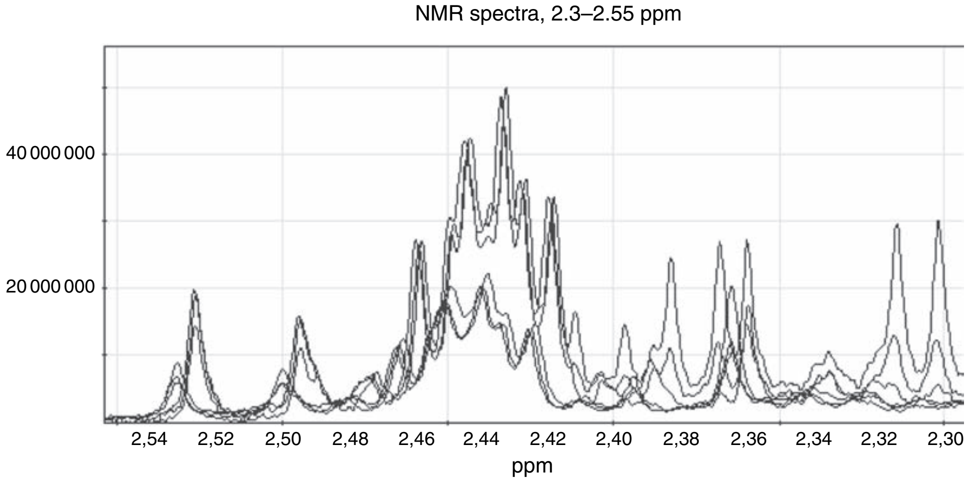


Figure 6.11. NMR spectra from both sampling periods clearly show the reason for the found separation in the score scatter plot.

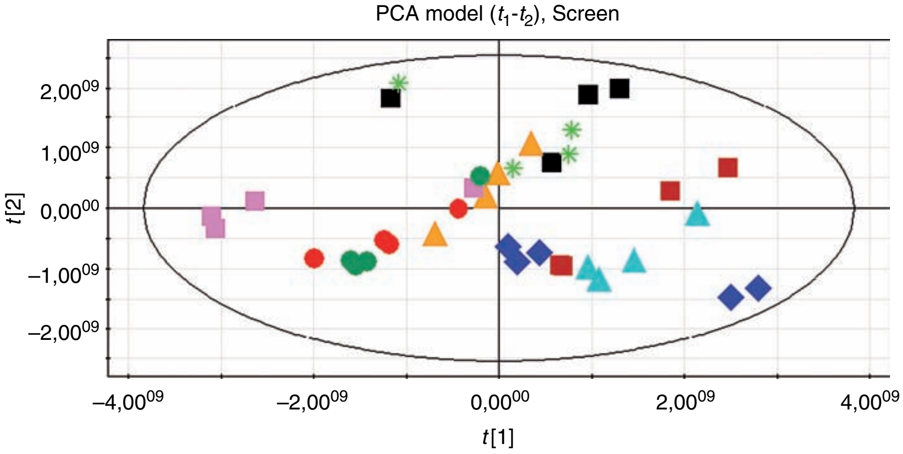


Figure 6.12. PCA score plot (t_1-t_2) after subtraction of the NMR screening sample from each individual. As shown in the plot this has corrected for the groupings due to different studies.

in modelling the effect of treatment for each individual over time. Again, a PCA model was calculated and its scores plot is found in Figure 6.12.

However, a greater inter-person than intra-person variation is still observed. Each colour corresponds to a different person.

The subtraction of the screening sample helped remove the separation between the sampling periods. However, there is still a problem in that the inter-person

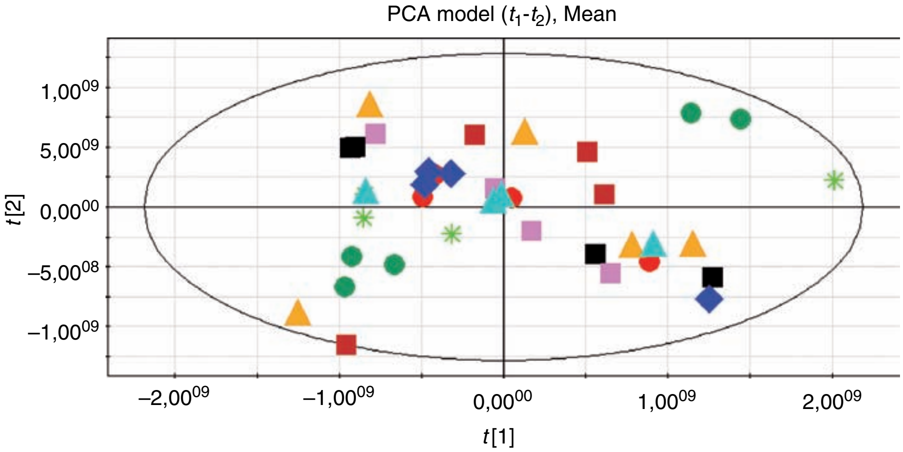


Figure 6.13. As shown in the PCA score plot (t_1-t_2), the systematic differences between objects and sampling periods have been removed by individual mean centring. Each colour corresponds to a different person.

variation is greater than the intra-person variation. This has an adverse influence on evaluating the effect of treatment over all objects. This is also an indication that the screening sample may not be a useful reference sample. One plausible reason may be due to the relatively long time period between the screening sample and the start of the study. It is important that the reference sample used is a biologically equivalent reference point for each object, if not, systematic differences between objects will exist. Hence, the average NMR spectrum for each object was used as their reference point, and the screening sample was excluded from further analysis.

The scores plot of the updated PCA model no longer displays any systematic differences between objects or sampling periods (see Figure 6.13).

However, it becomes clear that all individuals do not have the same behaviour over time following treatment. A number of reasons can exist, for example the absolute effect between individuals can be large, hence those with lower response will be suppressed in the model due to the scale-sensitivity of projection-based models such as PCA. Another reason can be that different individuals have different dynamic responses to treatment, for example quick and slow responders where the main effect for one individual occurs between time point 2 and 3, and for another person between time points 3 and 4.

One way to solve this problem is to create local or separate PCA models for each object, in order to identify the largest effect from start of sampling period (pre-dose). The assumption is that the largest change also reveals the largest effect of treatment. Hence, in the PCA scores plot for each object (individual) (shown in Figure 6.14), a direction of maximum change from the pre-dose sample is identified

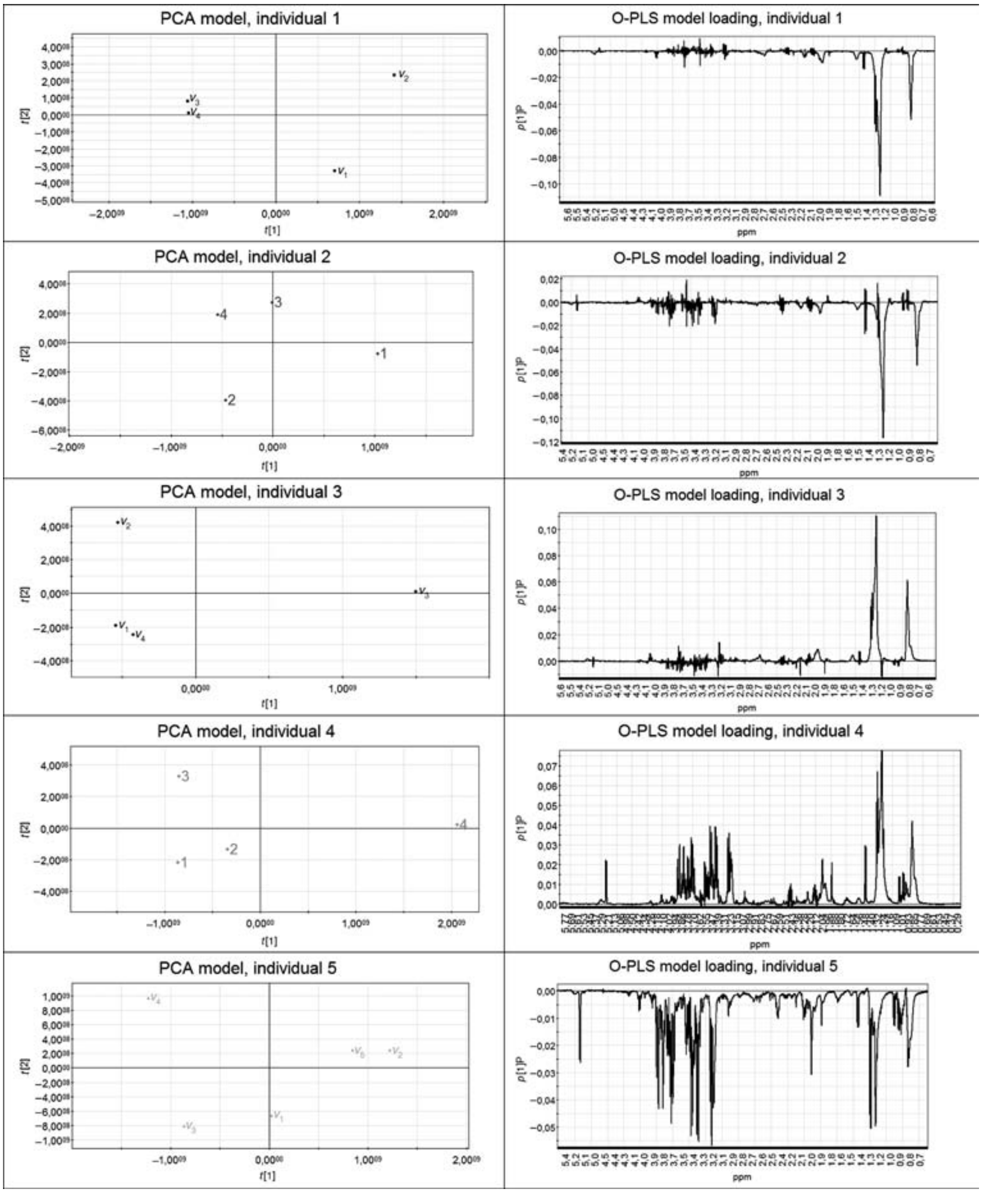


Figure 6.14. On the left, individual trajectories are shown, wherein the largest change over time can be identified. On the right, the OPLS loading profile of the predictive component for each individual is shown. (*Continued*)

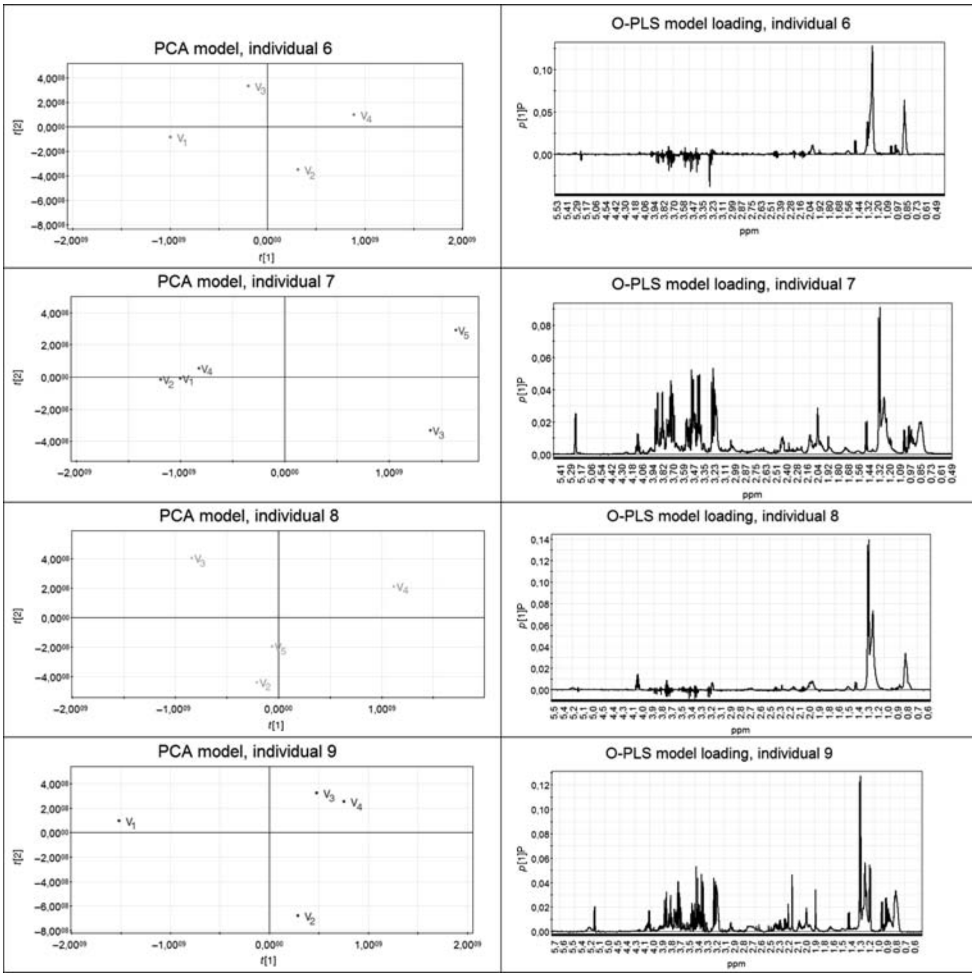


Figure 6.14. (Continued)

for one or several time points by assigning them with a discrete value of one (1), and all others, including the pre-dose sample with the value of zero (0). Following this, an OPLS-DA model was calculated in order to estimate the discriminating loading vector. This was repeated for all objects. The collective set of loading profiles are used to assign similarities between objects. For a summary of the loading vectors, see Figure 6.14.

A visual assessment of the OPLS-DA loading profiles, with lactate as the largest peak, shows two separate groups of profiles with opposite sign of the profile. Individuals 1, 2 and 5 represent one group of loading profiles and the others, individuals 3, 4, 6, 7, 8 and 9 make up for the second group. Here it should be strongly emphasized that one should make sure that this observed grouping is not due to the sampling periods, but rather to some underlying phenomenon.

In order to further validate the model, an OPLS-DA model was calculated, based on individuals 3 and 4 only. Those individuals have the most pronounced change in the PCA scores plot, and in addition, they also represent a slow and fast responder (maximum change is seen at different time points, see Figure 6.14). Individuals 6, 7, 8 and 9 were all excluded and used as an external prediction set. The observed vs predicted plot for the model and the prediction set is given in Figures 6.15, 6.16 respectively. The discriminate line (y-value of 0.25 given by the average of the y-vector used in the OPLS-DA model) means that only one sample is wrongly predicted!

It has to be emphasized that these groupings reflect the maximum change in the PCA model scores, and not necessarily the expected biological effect of treatment.

Early on in this study, four individuals were selected in agreement with a MVD for quantification of a few specific endogenous metabolites by HPLC analysis. Unfortunately, the most prominent metabolite in the loading profile, lactate, was not included as one of those metabolites. As a next step, OPLS modelling was performed with one of those endogenous metabolites as Y and their corresponding NMR measurements as X. Prior to modelling, the average of the endogenous metabolite for each person was removed in the Y-vector.

Individuals 5 and 6 were selected to establish a calibration model between the NMR-spectra and the quantitative measurements. A good model was obtained showing a fair correlation between the quantified concentrations of the metabolite and the calculated concentration (RMSEE = 0.19) see Figure 6.17.

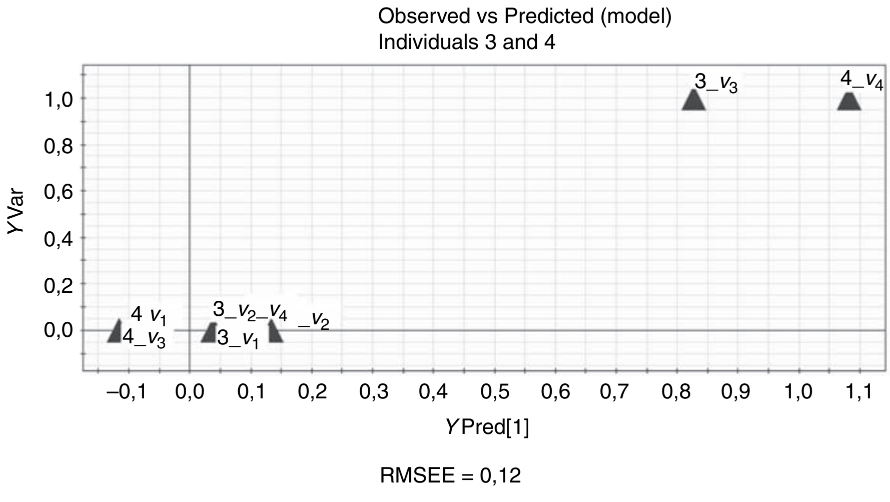


Figure 6.15. The OPLS-DA model shows a clear discrimination between samples having an observed effect to those where no effect was found.

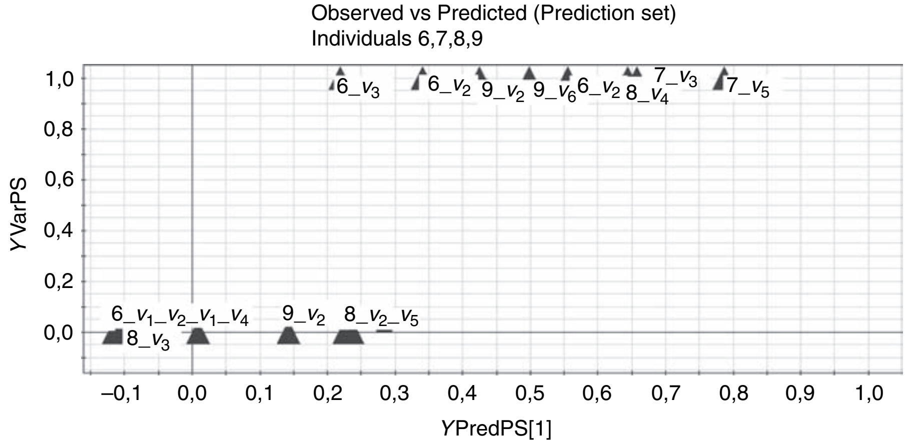


Figure 6.16. The OPLS-DA model predictions of an external test set resulted in only one wrongly predicted sample (6_v_3).

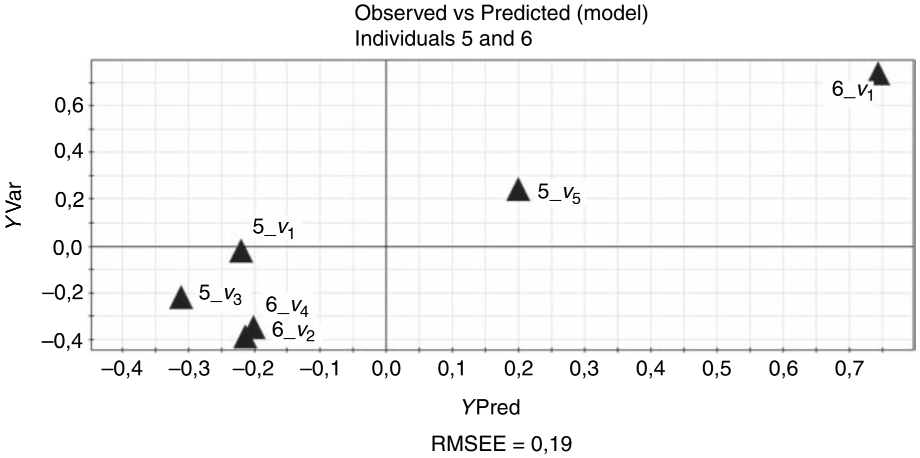


Figure 6.17. The quantified concentrations vs the calculated concentrations of the metabolite by the OPLS calibration model.

The calibration model was used for predictions of the metabolite concentration, for individual 1 and 8 at different time points, from the corresponding NMR-spectra. The predictions obtained vs the quantified metabolite concentration is shown in Figure 6.18. As shown in the figure the calibration model could be used for prediction of new samples with a good result. This indicates that we have identified relevant changes depending on treatment in at least one of the endogenous metabolites by using NMR-spectra.

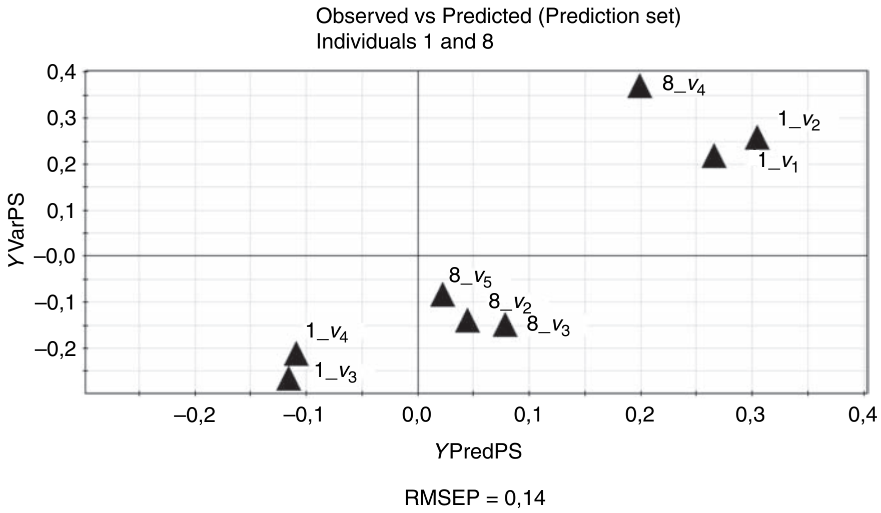


Figure 6.18. The OPLS model predictions of an external test set. The observed quantified concentrations are plotted vs the predicted concentrations of the endogenous metabolite.

6.3. Summary

In this chapter we have tried to guide the reader through a chemometrical approach for extracting information out of complex metabonomic data and this has been illustrated by an example. Wherein the most important findings are the usefulness of dynamic sampling, which provided us with an opportunity to identify slow, medium or fast responders as well as groups of objects showing different response profiles. By using this information all through the evaluation of the data, predictive models could be build on a small number of objects and finally validated by test sets. In the last part of the example we build a calibration model, wherein the NMR-profiles are used as the descriptor and a quantified metabolite was used as response. The endogenous metabolite was quantified by HPLC. This model was validated by an external test set.

The suggested chemometric approach to metabonomic studies is summarized in the following steps;

Step 1: Definition of aim
What is previously known?
What is needed to know?
How to reach those objectives?

Step 2: Study design
- Class specific studies
  - Objects or observations (e.g. samples) selected need to span the experimental domain, in a balanced and systematic manner
  - Apply multivariate design when selection of objects is possible
- Dynamic studies (investigate temporal progression)
  - Sampling over time

Step 3: Sample preparation and characterisation
- Experimental protocol/Analytical technique

Step 4: Evaluation of the collected data
- Data processing, for example GC-MS, LC-MS, LC-NMR
- Overview – PCA
- Classification/Discrimination – SIMCA method & OPLS-DA
  - One-class classifier (Control heterogenous)
  - Two-class classifier (Control class, Treated class)
- Prediction and biomarker identification

Finally we like to add a short list of questions that always should be asked of a data table independent of approach and/or methods used.

Questions about samples and observations (score plots)
- Are there any outliers?
- Are there groups and/or trends?
- Are there similarities/dissimilarities between samples?
- How do new samples behave?

Questions about variables (loading plots)
- Which variables cause outliers?
- Which variables are responsible for groupings and/or trends?
- Which variables are responsible for class separations?
- How do new variables behave?

6.4. Extensions and future outlook

Systems biology seeks to integrate information from multiple parts of a biological system in a holistic attempt to understand the whole system. A major concern is how to actually integrate multiple blocks of data, for example understand the relation between a data table *X* (e.g. a set of NMR spectral profiles) and another data table *Y* (e.g. a set of GC/MS resolved profiles). Current pattern recognition methods based on PLS methods, artificial neural networks, canonical correlation, support vector machines, and so on, all lack the proper model structure to describe these types of data structures, because they focus only on the *X-Y correlation overlap* and not on the *non-overlapping variation* (e.g. Y-orthogonal and X-orthogonal), which, in a biological sense, can be of equal interest. This is a fundamental problem as we certainly can not expect that all variation in NMR and GC/MS profiles co-vary. Here, the OPLS method and extensions thereof represent a good alternative.

6.4.1. Extensions of the OPLS model

The OPLS model structure can be extended to include X-orthogonal variation [36, 37]. The OPLS model then comprises of three sets of components representing

(i) the joint X–Y variation (given by the T_pP_p and U_pC_p components)
(ii) the Y-orthogonal ($T_oP_p{}^T$) variation
(iii) the X-orthogonal ($U_oC_o{}^T$) variation (Figure 6.19).

$$\begin{aligned} &\text{Model of } X\text{:} \qquad X = T_pP_p{}^T + T_oP_o{}^T + E \\ &\text{Model of } Y\text{:} \qquad Y = U_pC_p{}^T + U_oC_o{}^T + F \end{aligned}$$

E and F are the residual matrices of X and Y respectively. This can also be extended to more than two data tables, see for instance Reference [56], hence it nicely fits into a systems biology framework.

On the left, the Y-orthogonal components are shown, while on the right the X-orthogonal components are illustrated. In the middle section the predictive components are determined and the correlation between the two matrices are calculated.

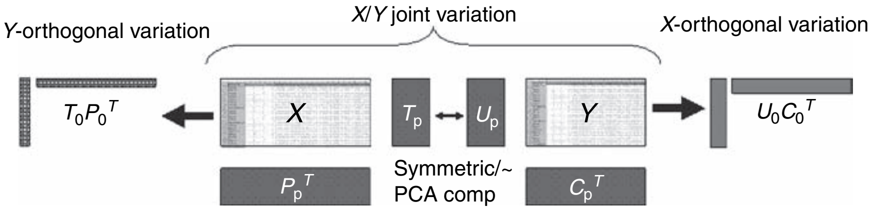


Figure 6.19. A graphical overview of the OPLS model where also the X-orthogonal variation is modelled. The Y-orthogonal variation ($T_oP_o{}^T$) represents the unique, non-overlapping variation in X, conversely, the X-orthogonal variation ($U_oC_o{}^T$) defines the unique systematic variation in Y. The X/Y joint variation (overlapping between X and Y) is given by the [$T_pP_p{}^T$, $U_pC_p{}^T$] components. This OPLS model structure is bi-directional, meaning that the model can be used for predictions in both directions.

6.4.2. Batch modelling

Batch modelling [26] is routinely being used for analysis of industrial batch process data. A batch process has a finite duration in time, in contrast to a continuous process. By analogy, batch modelling methods are used in metabonomic studies to model the time dependency or dynamics of biological processes, for example the evolution of the effects of a toxic substance in rats. Data collected from such studies produce a three-way data table where each dimensionality represents objects (e.g. rat urine or plant extract samples), variables (e.g. NMR shifts, m/z) and sample time points (see Figure 6.20). Batch modelling is based on modelling two levels, the observation level and the batch level. The observation level shows the dynamics of the biological process of each object over time, see Figure 6.16. For multiple objects (e.g. control rats), it is possible to establish an average trajectory with upper and lower limits based on standard deviations. These indicate the normal development of the object, for example control rats. The established control charts from the model can be used

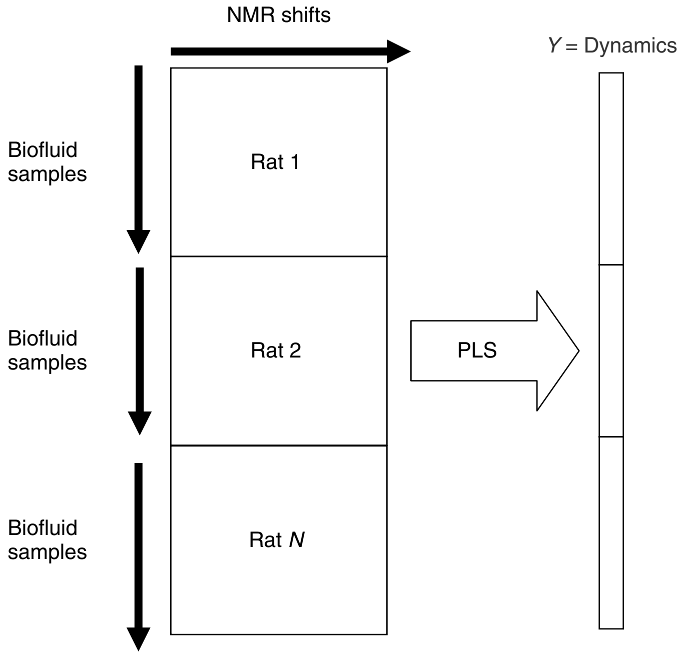


Figure 6.20. In batch modelling, the data is organized as an X-matrix containing blocks of rows where each block represents an object (e.g. a rat). Each row in a block represents the multivariate profile of an observation (e.g. the NMR spectral shifts) at a specific time point. The corresponding row in the Y-matrix contains the dynamics (e.g. the time point). This is followed by an PLS or OPLS model to extract variation from the X matrix related to the dynamics of the system.

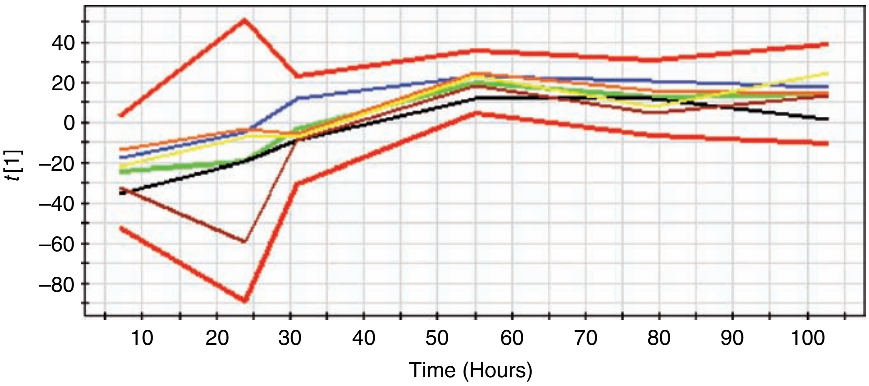


Figure 6.21. Batch control charts can be constructed from a PLS or OPLS batch model score vectors. The average score trajectory (for each component) with upper and lower control limits (based on standard deviations) indicates the normal dynamic trajectory for a batch. The control chart can be used for detecting deviations from normality.

to monitor the development of new objects and is used to detect deviations from normality, for example effect of a toxin or drug. Observed deviations from normality can be interpreted by means of contribution plots. Batch modelling is based on the assumption that a control group of objects is followed over the same time period as the treated group. Batch control charts can be constructed from a PLS or OPLS batch model score vectors. The average score trajectory (for each component) with upper and lower control limits (based on standard deviations) indicates the normal dynamic trajectory for a batch. The control chart can be used for detecting deviations from normality (see Figure 6.21).

6.4.3. Hierarchical PCA

The idea behind hierarchical PCA is to block the variables in order to improve transparency and interpretability [57–59]. This method operates on two or more levels, and on each level standard PCA scores and loading plots as well as residuals and their summaries such as DModX are used for interpretation. The procedure can, for two levels, be described as follows (see Figure 6.22). In the first step, in this case is to divide the large matrix into conceptually meaningful blocks and make a separate PCA for each matrix. In the next step the principal components (scores T) from each of these models become the new variables (“super variables”) describing the systematic variation from each block. In the final step a PCA model fitted to this data and the hierarchical PCA model is established, see Figure 6.22.

The interpretation of a hierarchical model has to be done in two steps. First, the loading plots of the hierarchical model reveal which of the blocks that are most

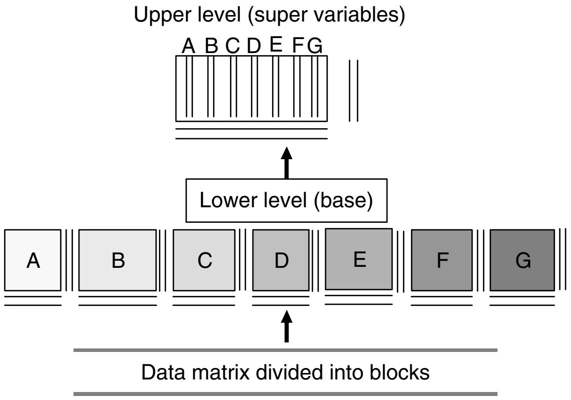

Upper level (super variables)
A B C D E F G
Lower level (base)
A
B
C
D
E
F
G
Data matrix divided into blocks

Figure 6.22. H-PCA is shown from the bottom to the top. At the bottom of the figure, the data matrix is divided into blocks. A separate PCA model is calculated for each block and the PCA score components from each model are then combined to form a new matrix, summarising all blocks. This new block of data is then analyzed by a PCA.

important for any groupings that can be seen in the hierarchical score plot. Second, the loading plots for the blocks of interest are studied on the lower level and in the corresponding loading plot the original variables of importance can be identified. The Hierarchical PCA is easily extended to one type of hierarchical PLS or PLS/DA by adding a *Y* (response/discriminate) matrix on the upper level. The interpretation is done in analogy with PLS or PLS/DA on the upper level and as in H-PCA on the lower level.

References

[1] Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Michotte, Y., Kaufman, L., *Chemometrics: A textbook*, Elsevier, 1988.

[2] Martens, H., Naes, T., *Multivariate Calibration*, Wiley: Chichester, 1989.

[3] Eriksson, L., Johansson, E., Kettaneh Wold, N., Wold, S., *Multi and Megavariate Data Analysis*, Umetrics, 2001.

[4] Gullberg, J., Jonsson, P., Nordström, A., Sjöström, M., Moritz, T., Design of Experiments: An Efficient Strategy to Identify Factors Influencing Extraction and Derivatization of Arabidopsis Thaliana Samples in Metabolomic Studies with Gas Chromatography/Mass Spectrometry. *Analytical Biochemistry* 331, 283–295, 2004.

[5] Jiye, A., Trygg, J., Gullberg, J., Johansson, A.I., Jonsson, P., Antti, H., Marklund, S.L., Moritz, T., Extraction and GC/MS Analysis of the Human Blood Plasma Metabolome, *Analytical Chemistry* 77: 8086–8094, 2005.

[6] Idborg-Björkman, H., Edlund, P.O., Kvalheim, O.M., Schuppe-Koistinen, I., Jacobsson, S.P., *Analytical Chemistry* 75: 4784–4792, 2003.

[7] Potts, B.C.M., Deese, A.J., Stevens, G.J., Reily, M.D., Robertson, D.G., Theiss, J., NMR of Biofluids and Pattern Recognition: Assessing the Impact of NMR Parameters on the Principal Component Analysis of Urine from Rat and Mouse. *Journal of Pharmaceutical and Biomedical Analysis* 26: 463–476, 2001.

[8] Robertson, D.G., Reily, M.D., Sigler, R.E., Wells, D.F., Paterson, D.A., Braden, T.K., Metabonomics: Evaluation of Nuclear Magnetic Resonance (NMR) and Pattern RecognitionTechnology for Rapid in Vivo Screening of Liver and Kidney Toxicants, *Toxicological Sciences* 57: 326–337, 2000.

[9] Holmes, E., Antti, A., Chemometric Contributions to the Evolution of Metabonomics: Mathematical Solutions to Characterising and Interpreting Complex Biological NMR Spectra. *Analyst* 127: 1549–1557, 2002.

[10] Nicholson, J.K, Connelly, J., Lindon, J.C., Holmes, E. Metabonomics: A Platform for Studying Drug Toxicity and Gene Function. *Nature Reviews Drug Discovery* 1: 153–161, 2002.

[11] Jackson, J.E., *A Users Guide to Principal Components*. Wiley, New York, 1991.

[12] Wold, S., Ruhe, A., Wold, H., Dunn III, W.J. The Collinearity Problem in Linear Regression. The Partial Least Squares Approach to Generalized Inverses. *SIAM J. Sci. Stat. Comput.* 5: 735–743, 1984.

[13] Wold, S. Martens, H. Wold, H., Lecture Notes in Mathematics, Proc. Conf. Matrix pencils, Piteå, Sweden, Springer Verlag, Heidelberg, 1983.

[14] Wold, S., Eriksson, L., Sjöström, M. PLS in Chemistry, Schleyer P.V.R (ed.) *Encyclopedia of Computational Chemistry*, John Wiley & Sons, New York, 2006–2016, 1998.

[15] Wold, S., Albano, C., Dunn III, W.J., Edlund, U., Esbensen, K., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W., Sjöström, M., *Multivariate Data Analysis in Chemistry*, NATO ASI Series C 138, D. Reidel Publ. Co., Dordrecht, Holland, 1984.

[16] Lundstedt, T., Seifert, E., Abramo, L., Thelin, B., Nyström, A., Pettersen, J., Bergman, R. Experimental Design and Optimization. *Chem Intel Lab Systems* 42: 3–40, 1998

[17] Box, G.E.P., Hunter, W.G., Hunter, J.S. *Statistics for Experimenters*, John Wiley & Sons, New York, 1978.

[18] Eriksson, L., Johansson, E., Kettaneh Wold, N., Wikström, C., Wold, S. *Design of Experiments – Principles and Applications*, Umetrics AB, 1996.

[19] Gemperline, P.J. Developments in Nonlinear Multivariate Calibration. *Chemometrics and Intelligent Laboratory Systems* 15: 115–126, 1992.

[20] Kowalski, B.R., Seasholtz, M.B., Recent Developments in Multivariate Calibration. *Journal of Chemometrics* 5: 129–145, 1991.

[21] Hellberg, S., Sjöström, M., Skagerberg, B., Wold, S. Peptide Quantitative Structure-Activity-Relationships, a Multivariate Approach. *Journal of Medicinal Chemistry* 30 : 1126–1135, 1987.

[22] Lundstedt, T., A QSAR strategy for screening of drugs – and predicting their clinical activity. *Drugs, News & Perspectives* 4(8) 468, 1991.

[23] Wold, S., Albano, C., Dunn III, W.J., Esbensen, K., Hellberg, S., Johansson, E., Sjöström, M. Pattern Recognition: Finding and Using Regularities in Multi-Variate Data, in Food Research and Data Analysis (eds H.M. Russwarm Jr H.) 147–188 Applied Science Publishers, London, 1983.

[24] Albano, C., Dunn III, W.J., Edlund, U., Johansson, E., Nordén, B., Sjöström, M., Wold, S. Four Levels of Pattern Recognition. *Analytica Chimica Acta* 103: 429–443, 1978.

[25] Wold, S., Pattern Recognition by Means of Disjoint Principal Components Models. *Pattern Recognition* 8, 127–139, 1976.

[26] Wold, S., Kettaneh, N., Friden, H., Holmberg, A. Modelling and Diagnostics of Batch Processes and Analogous Kinetic Experiments. *Chemometrics and Intelligent Laboratory Systems* 44: 331–340, 1998.

[27] Kourti, T., MacGregor, J.F., Multivariate SPC Methods for Process and Product Monitoring. *Journal of Quality Technology* 28: 409–428, 1996.

[28] Workman, J., Veltkamp, D.J., Burgess, L.W., Process Analytical Chemistry. *Analytical Chemistry* 71: 121R–180R, 1999.

[29] Hotelling, H. The Most Predictable Criterion. *Journal of Educational Psychology* 26: 139–142, 1935.

[30] Greenacre, M.J., *Theory and Applications of Correspondence Analysis*. London: Academic Press, 1984.

[31] Bishop, C.M., *Neural Networks for Pattern Recognition*, Oxford, 1996.

[32] Wythoff, B.J., Backpropagation Neural Networks – A Tutorial, *Chemometrics and Intelligent Laboratory Systems* 18(2) 115–155, 1993.

[33] Sivia, D.S. *Data Analysis: A Bayesian Tutorial*. Oxford: Oxford University Press, 1996.

[34] Rabiner, L.R., Juang, B.H., An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, January, 1986.

[35] Trygg, J., Wold, S., Orthogonal Projections to Latent Structures (O-PLS). *Journal of Chemometrics* 16: 119–128, 2002.

[36] Trygg, J., O2-PLS for Qualitative and Quantitative Analysis in Multivariate Calibration. *Journal of Chemometrics* 16: 283–293, 2002.

[37] Trygg, J., Wold, S., O2-PLS, A Two-block (*X-Y*) Latent Variable Regression (LVR) Method with an Integral OSC Filter, *Journal of Chemometrics* 17: 53–64, 2003.

[38] Cloarec, O., Dumas, M.E., Trygg, J., Craig, A., Barton, R.H., Lindon, J.C., Nicholson, J.K., Holmes, E., Evaluation of the Orthogonal Projection on Latent Structure Model Limitations Caused by Chemical Shift Variability and Improved Visualization of Biomarker Changes in H-1 NMR Spectroscopic Metabonomic Studies, *Analytical Chemistry* 77(2): 517–526, Jan. 15, 2005.

[39] Kvalheim, O.M., The Latent Variable. *Chemometrics Intelligent Laboratory systems* 14: 1–3, 1992.

[40] Wold, S., Sjöström, M., Carlson, R., Lundstedt, T., Hellberg, S., Skageberg, B., Wikström, C., Multivariate Design. *Analytica Chimica Acta* 191: 17, 1986.

[41] Carlson, R., Lundstedt, T., Scope of Organic Synthetic Reactions. Multivariate Methods for Exploring the Reactin Space. An Example of the Willgerodt-Kindler Reaction. *Acta Chemica Scandinavia* B 41: 164, 1987.

[42] Carlson, R., Lundstedt, T., Albano, C., Screening of Suitable Solvents for Organic Synthesis, Strategies for Solvent Selection. *Acta Chemica Scandinavica* B 39: 79, 1984.

[43] Sandberg, M., Sjöström, M., Jonsson, J., A Multivariate Characterization of tRNA Nucleosides. *Journal of Chemometrics* 10: 493–508, 1996.

[44] Oprea, T.I., Gottfries, J., Chemography: The Art of Navigating in Chemical Space. *Journal of Combinatorial Chemistry* 3(2), 157–166, Mar–Apr 2001.

[45] deAguiar, P.F., Bourguignon, B., Khots, M., Massart, D.L., PhanThanLuu, R., D-optimal Designs. *Chemometrics and Intelligent Laboratory Systems* 30(2), 199–210, 1995.

[46] The Standard Metabolic Reporting Structure, Version 2.3, http://www.smrsgroup.org/, January 13 2006.

[47] Jonsson, P., Gullberg, J., Nordström, A., Kowalczyk, M., Sjöström, M., Moritz, T., A Strategy for Extracting Information from Large Series of Non-processed Complex GC/MS Data. *Anal Chem.* 76: 1738–1745, 2004.

[48] Jonsson, P., Bruce, S.J., Moritz, T., Trygg, J., Sjöström, M., Plumb, R., Granger J, Maibaum, E., Nicholson, J.K., Holmes, E., Antti, H., Extraction, Interpretation and Validation of Information for Comparing Samples in Metabolic LC/MS Data Sets. *Analyst* 130: 701–707, 2005.

[49] Halket, J.M., Przyborowska, A., Stein, S.E., Mallard, W.G., Down, S., Chalmers, R.A. Deconvolution Gas Chromatography Mass Spectrometry of Urinary Organic Acids – Potential

for Pattern Recognition and Automated Identification of Metabolic Disorders. *Rapid Commun. Mass Spectrom.* 13: 279–284, 1999.

[50] Shen, H.L., Grung, B., Kvalheim, O.M., Eide, I., Automated curve resolution applied to data from Multi-detection Instruments. *Analytica Chimica Acta* 446 (1–2), 313–328, Nov. 19 2001.

[51] Torgrip, R.J.O., Aberg, M., Karlberg, B., Jacobsson, S.P., Peak Alignment Using Reduced Set Mapping, *Journal of Chemometrics* 17(11), 573–582, Nov. 2003.

[52] Vogels, J.T.W.E., Tas, A.C., van den Berg, F., van der Greef, J., A New Method for Classification of Wines Based on Proton and Carbon-13 NMR Spectroscopy in Combination with Pattern Recognition Techniques. *Chemometrics and Intelligent Laboratory Systems* 21, 2–3, 249–258, 1993.

[53] Hotelling, H., The Generalization of Student's Ratio. *Ann. Math. Statist.* 2, 360–378, 1931.

[54] Miller, P., Swanson, R.E., Heckler, C.E., Contribution Plots: A Missing Link in Multivariate Quality Control. *Appl. Math. And Comp. Sci.* 8(4), 775–792, 1998.

[55] Cloarec, O., Dumas, M., Craig, E., Barton, R.H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J.C., Holmes, E., Nicholson, J., *Analytical Chemistry* 77, 1282–1289, 2005.

[56] Eriksson, L., Damborsky, J., Earll, M., Johansson, E., Trygg, J., Wold, S., Three-block Bi-focal PLS (3BIF-PLS) and Its Application in QSAR, SAR QSAR Environmental Research, 15(5–6) 481–499, Oct.–Dec. 2004.

[57] Wold, S., Kettaneh, N., Tjessem, K., Hierarchical Multiblock PLS and PC Models for Easier Model Interpretation and as an Alternative to Variable Selection, *Journal of Chemometrics* 10(5–6), 463–482, Sep.–Dec. 1996.

[58] Eriksson, L., Johansson, E., Lindgren, F., Sjöström, M., Wold, S., Megavariate Analysis of Hierarchical QSAR Data. *Journal of Computer-Aided Molecular Design* 16: 711–726, 2002.

[59] Gunnarsson, I., Andersson, P.M., Wikberg, J., Lundstedt, T., Multivariate Analysis of G Protein-coupled Receptors, *Journal of Chemometrics* 17: 82–92, 2003.