Chapter 7
Chemometrics Methods and Strategies in Metabolomics

Rui Climaco Pinto

Abstract Chemometrics has been a fundamental discipline for the development of metabolomics, while symbiotically growing with it. From design of experiments, through data processing, to data analysis, chemometrics tools are used to design, process, visualize, explore and analyse metabolomics data.

In this chapter, the most commonly used chemometrics methods for data analysis and interpretation of metabolomics experiments will be presented, with focus on multivariate analysis. These are projection-based linear methods, like principal component analysis (PCA) and orthogonal projection to latent structures (OPLS), which facilitate interpretation of the causes behind the observed sample trends, correlation with outcomes or group discrimination analysis. Validation procedures for multivariate methods will be presented and discussed.

Univariate analysis is briefly discussed in the context of correlation-based linear regression methods to find associations to outcomes or in analysis of variance-based and logistic regression methods for class discrimination. These methods rely on frequentist statistics, with the determination of *p*-values and corresponding multiple correction procedures.

Several strategies of design-analysis of metabolomics experiments will be discussed, in order to guide the reader through different setups, adopted to better address some experimental issues and to better test the scientific hypotheses.

Keywords Metabolomics • Data analysis • Chemometrics • Multivariate • Exploratory • Regression • Classification • Discrimination • Discovery • Validation

R.C. Pinto
Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London,
St. Mary's Campus, Norfolk Place, W2 1PG, London, England, UK
e-mail: r.pinto@imperial.ac.uk

A. Sussulini (ed.), *Metabolomics: From Fundamentals to Clinical Applications*,
Advances in Experimental Medicine and Biology,
DOI 10.1007/978-3-319-47656-8_7

Abbreviations

ANOVA	Analysis of variance
ASCA	ANOVA-simultaneous component analysis
AUC	Area under the curve (in the context of ROC curves)
CV	Cross-validation
CV-ANOVA	Cross-validation – analysis of variance
FWER	Family-wise error rate
FDR	False discovery rate
GC-MS	Gas chromatography coupled to mass spectrometry
HCA	Hierarchical cluster analysis
ICA	Independent component analysis
iQC	Internal quality control (sample)
IS	Internal standard
LC-MS	Liquid chromatography coupled to mass spectrometry
LOO	Leave-one-out procedure in cross validation
MS	Mass spectrometry
MWAS	Metabolome-wide association studies
MWSL	Metabolome-wide significance level
OPLS	Orthogonal projections to latent structures
OPLS-DA	Orthogonal projections to latent structures – discriminant analysis
OPLS-EP	Orthogonal projections to latent structures – effect projection
PC	Principal component
PCA	Principal component analysis
PLS	Projections to latent structures
PRESS	Predicted residual error sum of squares
R2X	Fraction of variance in the data explained by each latent variable
R2Y	Fraction of variance of y/Y explained by each latent variable
ROC	Receiver operating characteristic (curve)
Q2	Model statistics to evaluate quality of model prediction
RMSECV	Root mean squared error of cross validation
RMSEP	Root mean squared error of prediction
SMART	Scaled-to-maximum, aligned and reduced trajectories
SUS	Shared and unique structures
VIP	Variable importance on projection

7.1 Introduction

Metabonomics [1] or metabolomics [2] concerns the study of the metabolome, a multivariate ensemble of small molecules that are intermediates and products of metabolism. Its main emphasis is on metabolite profiling, at the level of cells or organs, of endogenous and/or exogenous metabolites, and on the effects of perturbations of the metabolism caused by disease, environmental, or dietary influences.

Chemometrics can be defined as “the chemical discipline that uses mathematical, statistical, and other methods employing formal logic, to design or select optimal measurement procedures and experiments, and to provide maximum relevant chemical information by analysing chemical data” [3]. It differs from statistics in analytical chemistry mainly due to its computer intensive nature, and for being mostly multivariate analysis based [4]. Due to the nature of the signals in chemistry, namely in spectroscopy, chemometrics developed around the subject of multivariate analysis, because of its ease of interpretation. These are correlation-/projection-based methods, which require computationally intensive work. While bioinformatics and chemoinformatics are also used for data analysis, they are more related to data mining and use of databases. These disciplines have some overlap with chemometrics and methods like principal component analysis (PCA), for instance, are used by all of them.

Chemometrics is intensively used in the metabolomics context due to its experimental design component and to the fact that metabolic systems are multivariate in nature, with data mostly a product of ^{1}H NMR spectroscopy and gas/liquid chromatography coupled to mass spectrometry (GC/LC-MS). Metabolomics naturally relates to clinical research due to the fact that specific metabolite profiles express themselves in a living organism through a resulting health phenotype.

Clinical experiments exist in different areas and contexts, such as understanding biological processes and disease mechanisms, in vitro studies of materials of human origin, models of human disease processes, follow-up after surgery, epidemiological studies, diagnostic and therapeutic methods, effect and mechanism of vaccines and drugs, biomarker discovery and disease discrimination, among others [5]. Chemometrics may help unravel information from metabolomics in different aspects of each of these contexts. Although not specifically designed for clinical research, the methods presented in this chapter adapt to the field naturally, as they can be used to explore clinical metabolomics data.

This chapter is devoted to the uses of state-of-the-art chemometrics methods and their application to metabolomics data in clinical analysis.

7.2 Notation

Notation in the text is as follows: vectors are presented in bold lower-case (e.g. **y**), matrices in bold upper-case (e.g. **X**) and indexes in italic lower-case letters (e.g. *i*). The metabolite data matrix **X** consists of samples in *i* rows and metabolic features (or metabolites) in *j* columns. Each continuous or discrete outcome **y** (e.g. blood pressure) has the same length *i* as rows in **X**. To define classes for the two-class case, a dummy vector **y** (e.g. 0 = control; 1 = disease) is built. In case there are more than two classes, a dummy matrix **Y** with one vector per class is built. Confounder factors, when mentioned, are vectors **z** with the same length as the rows in **X**. Qualitative confounder vectors are transformed into dummy matrices the same way as described for multiple classes. In case there are two or more confounders, they are horizontally concatenated into a matrix **Z**. Transposed matrix is indicated by using the letter “T” in superscript, as in $\mathbf{X}^{\mathrm{T}}$.

7.3 Data Preprocessing

While using univariate analysis, there is no need for variable normalization (unless normal distribution is deemed necessary) because each metabolic feature is evaluated separately; however, in multivariate analysis, normalization is of utmost importance and depends on the analysis in question. As preprocessing, normalization, scaling and transformations of data are discussed in Chap. 6 of this book, they will not be herein discussed in detail. We assume the samples were already normalized with the objective of reducing magnitude effects (e.g. caused by different dilution levels), and the variables were scaled in an appropriate way (e.g. ^{1}H NMR was Pareto scaled; LC-MS was centred and unit variance scaled) and potentially transformed adequately (log transformation or other). Both ^{1}H NMR and MS data are now considered a data matrix **X** of metabolic features ready for statistical analysis.

7.4 Chemometrics Contexts and Methods

The need for chemometrics tools arises around three decades ago, due to the development of more complex instruments with a consequent increase in the number of variables, and is propelled by the development of computational capacity. Large-scale dataset simultaneous visualization is more difficult in a univariate approach, and, for example, multiple regression modelling is constrained by variable colinearity. As referred previously, the chemometrics discipline is based on computing intensive methods, in general multivariate, which solves the colinearity problem in a covariance-/correlation-based framework.

There are many different multivariate methods for modelling data, as shown in previous literature reviews [6–8]. They can be unsupervised (no assumptions made on the samples) or supervised (samples are defined into classes, or each sample is associated to an outcome y_i value). Multivariate methods represent the samples as points in the space of the initial variables. The samples can then be projected into a lower dimensionality space – into components or latent variables – such as a line, a plane or a hyperplane, which can be seen as the “shadow” of the dataset viewed from its “best” viewpoint. The coordinates of the samples in the newly defined latent variables are defined as the scores, while the directions of variance to which they are projected are defined as the loadings. The loadings vector for each latent variable contains the weights of each of the initial variables in that latent variable. For a certain latent variable, the more a sample score is distant from its centre, the higher values it has in some of the initial variables (while potentially having lower values in others). Respectively, these initial variables have high weights in the loadings vector of that latent variable.

Projection-based linear methods are popular due to the simplicity of interpretation, thus used when understanding of a system is important. Nonlinear methods such as neural networks, support vector machines and random forests are less common in metabolomics when interpretation is needed, and are used mostly for prediction of new samples in classification/regression contexts.

At the moment, due to the large amount of features involved in untargeted metabolomics, most of the statistical methods are applied previously to compound/metabolite identification. Only after finding a smaller number of important statistically significant metabolic features (putative metabolites), the analyst proceeds to the identification phase, as this may be very time-consuming. Bayesian networks have also been recently used in metabolomics but are not purely based on numerical metabolomics data. Because of their need for extra information, including metabolite identification and/or information from databases, these methods are considered to be more in the bioinformatics than in the chemometrics domain; thus, they will not be discussed here.

7.4.1 Multivariate Data Exploration (PCA)

The simplest correlation- and projection-based multivariate analysis linear method, and simultaneously the most widely used tool in chemometrics, is principal component analysis (PCA) [9–12]. It can be seen as the basis for other multivariate methods, thus being commonly used to introduce the concept of latent variables, and it is widely used as an exploration tool in metabolomics [13].

PCA is a non-supervised method. As it contains no assumptions on the data, it is used as a visualization and exploration tool at the start of any analysis, in order to detect trends, groups and outliers. It allows simpler global visualization by representing the variance in a small number of uncorrelated latent variables, which can then be understood to be information or random variation.

PCA decomposes the data matrix into principal components (latent variables or latent structures) that represent the underlying structure of the data. This allows one to represent the structured variance in the data by a smaller number of (latent) variables, while discarding the noise, thus making it appropriate for dimensional reduction. A matrix **X** (of e.g. metabolites) is decomposed by PCA using p components as follows: $\mathbf{X} = \mathbf{T}.\mathbf{P}^{\mathrm{T}} + \mathbf{E}$, where **X** has dimensions $n \times m$, **T** is a $n \times p$ matrix of scores, **P** is a $m \times p$ matrix of loadings and **E** is a $n \times m$ matrix with residual variance, i.e. not included in the latent variable model. Depending on the objective of the analysis, the number of components in the model can be decided arbitrarily (e.g. a number "large enough"), according to a certain percentage of variance described with that number of components (e.g. 95 % of cumulative variance), or by using cross-validation strategies (which are later described).

An example of a PCA analysis is depicted in Fig. 7.1. Scores are coloured according to some meta-information after PCA calculations, in order to understand the reasons for the clusters. Samples in the same cluster are similar in the components represented, while variables in the same clusters are correlated with each other. To see, e.g. which variables are higher/lower in group B, draw a line passing in the centre of group B and through zero and then draw a line in the same direction

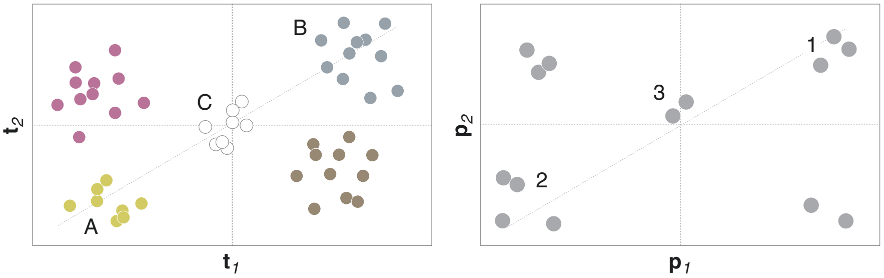

Fig. 7.1 PCA scatter plots of scores $\mathbf{t}_1$ vs $\mathbf{t}_2$ (*left*) and loadings $\mathbf{p}_1$ vs $\mathbf{p}_2$ (*right*) should be inspected simultaneously in order to understand the relations between trends and groups observed in the samples (score plot) and which variables – metabolites – are responsible for it (loadings plot)

in the loadings plot. Variables 1 are over that line in the same area as samples from group B; thus they are in higher values in B than in, e.g. group A (which is on the opposite side). Inversely, variables 2 have higher values in group A. Samples C, located close to the origin, have average behaviour between A and B. Variables 3 have no influence in this component, as their weights in the loadings of PC1 and PC2 are close to zero. Note that PC1 vs PC2 are being shown, but due to PCA's orthogonality of components, any PC can be plotted perpendicular to each other. In addition, sometimes plots of three components (xyz) are used, although they may become too complex to visualize due to the number of features involved.

Mathematically, the first principal component is the line that better approximates the data, in the least-squares sense. It represents thus the direction of the largest variance in the dataset, or in other words, the direction in which the variance of the coordinates of the samples is maximized. The dataset information explained by the first component can be subtracted from the initial data, and a second component can then be calculated from the residuals. Each principal component (PC) represents a fraction of the variance in the data – a pattern that can be in higher or lower magnitude in each sample – and is unrelated (orthogonal in a linear algebra sense, perpendicular in a geometrical sense) to the others (thus can be drawn perpendicularly to each and every other). The orthogonality property of PCA can be easily understood if one considers the calculation of each PC at a time. After PC_i is calculated from a data matrix $\mathbf{X}$, the information it represents is deleted from $\mathbf{X}$. Thus, for the calculation of the next component PC_{i+1}, that information is not available anymore.

Apart from helping at visualizing trends and groups in the data, an important application of PCA is to look for outliers in the samples. Outliers are samples that have scores very distant (thus different) from the others. They can be found by inspecting the scores or a model's cumulative measure of distance such as Hotelling T^2 [14], as well as by inspecting the residuals of the model (large residuals may indicate mild outliers). Due to their high leverage during model creation, special care must be taken in order to remove them or not, prior to defining a model and interpreting it. It may make sense to remove outliers, if one understands they are caused by gross errors during sample preparation or instrumental analysis. More difficult decisions arise for less extreme samples, in which the large score distance to other samples cannot be justified by that, but is the result of correctly measured high or low values in some variables. Many different ways exist to look for multivariate outliers [15–18]. Robust algorithms, which can better at handling outliers, have been developed for PCA [19]. Note: an extensive literature list on PCA can be found on http://www.stats.org.uk/pca.

7.4.2 Multivariate Regression (OPLS)

Projection to latent structures (PLS) [20] is a supervised multivariate linear regression method similar in concept to PCA, which finds the relations between two matrices (data **X** and response **Y**), by maximizing the covariance of their latent variables. It allows to understand which variables (e.g. metabolites) of **X** are more correlated to the response (e.g. calcium levels in blood) and to make predictions for new samples.

Orthogonal projection to latent structures (OPLS) [21] is a modification of the PLS method. OPLS has the same predictive power as PLS but provides better interpretation of the relevant variables than PLS. It does so by decomposing the data in so-called "predictive" information related to the response **Y** (as concentrations, classes), "orthogonal" structured information not related to the response (as instrumental, biological variations) and residual variation.

The decomposition of a matrix **X** by OPLS for the single-*y* case using *p* latent variables is as follows [22]: $\mathbf{X}=1.\bar{x}^{\mathrm{T}}+\mathbf{t}_{\mathrm{p}}.\mathbf{p}_{\mathrm{p}}{}^{\mathrm{T}}+\mathbf{T}_{\mathrm{o}}.\mathbf{P}_{\mathrm{o}}{}^{\mathrm{T}}+\mathbf{E}$, where the data matrix **X** has dimensions $n\times \mathrm{m}$, 1 is a vector of dimension $n\times 1$ with ones in all positions, $\bar{x}$ is a vector $n\times 1$ with the column averages of **X**, $\mathbf{t}_{\mathrm{p}}$ is a vector of $n\times 1$ predictive scores, $\mathbf{p}_{\mathrm{p}}$ is a vector of $n\times 1$ predictive loadings, $\mathbf{T}_{\mathrm{o}}$ is a $n\times p-1$ matrix of orthogonal scores, $\mathbf{P}_{\mathrm{o}}$ is a $m\times p-1$ matrix of orthogonal loadings and **E** is a $n\times m$ matrix with residual variance, not included in the latent variable model, as it contains only residual, nonstructured variation.

The model prediction of a **y** variable by OPLS is obtained by $\mathbf{y}=\bar{\mathrm{y}}+\mathbf{t}_{\mathrm{p}}.\mathbf{q}_{\mathrm{p}}{}^{\mathrm{T}}+\mathbf{r}$, in which **y** is a response vector of dimensions $n\times 1$, $\bar{\mathrm{y}}$ is a vector of dimension $n\times 1$ with the average of **y** in all positions, **t** is the predictive scores vector from **X** and **q** is a vector $n\times 1$of predictive loadings from **y**, while **r** is a $n\times 1$ vector of **y** residuals.

Notice that for the single-y case, there can be only one predictive component, although many orthogonal ones may exist. Because of the predictive and orthogonal variance decomposition, one can look at the predictive score direction from negative to positive as an increase in the magnitude of **y**, which is positively correlated with variables in the positive side of the predictive loadings (and inversely correlated with variables on the negative side). For the multiple-y case, there may be multiple predictive components, reflecting the overlap in information between the matrices **X** and **Y**. Figure 7.2 illustrates single-y OPLS analysis.

OPLS is the multivariate linear method of choice to, e.g. find metabolic biomarkers correlated with a continuous variable, such as calcium score or blood pressure.

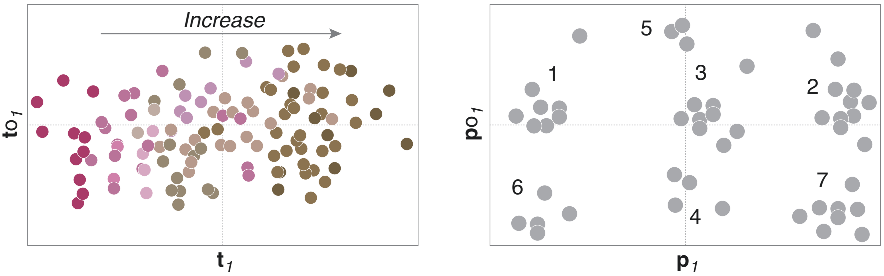

Fig. 7.2 Single-y OPLS scatter plots of scores (*left*) and loadings (*right*) for predictive component 1 vs orthogonal component 1, with scores coloured by **y** variable (e.g. blood pressure). OPLS models the **y** variable in the predictive component; thus samples with positive score t_1 (*right side* of the scores plot) are more concentrated on variables on the positive side of $\mathbf{p}_1$ (clusters 2 and 7) and less concentrated in variables with negative $\mathbf{p}_1$ (clusters 1 and 6). The orthogonal variation that is seen in the orthogonal scores to_1 (up–down) can also be inspected by colouring the scores according to different meta-information (e.g. gender, age) or the loadings (e.g. compound class). Variables related to a trend in the orthogonal scores are found along the orthogonal component loadings po_1

7.4.3 Multivariate Classification/Class Discrimination (OPLS-DA)

OPLS discriminant analysis (OPLS-DA) [23] has been largely used in the metabolomics context, and it is now the multivariate linear model of choice for classification/discrimination [24]. The term classification is used when the objective is to classify new objects into one of two or more possible classes (e.g. control, disease A, disease B). The term discrimination is used for the two-class case, in which the objective is to separate two classes and investigate the causes for class separation (e.g. biomarker discovery or which metabolites are in higher/lower concentration in a disease class in relation to a control class). Figure 7.3 shows an OPLS-DA example.

Notice that in OPLS the vector *y* is a continuous variable; in two-class discrimination, OPLS-DA **y** is categorical and, thus, defined as a dummy vector of 0/1 for the two-class case (for the multiple-y case, it is a dummy matrix with a 0/1 vector per class), describing class belonging. Although multi-class OPLS-DA can be calculated, most of the applications in metabolomics use a two-class model, as the interpretation is much more straightforward. Strategies for multiple class comparison using OPLS-DA are presented later in the chapter.

7.4.4 Note on Orthogonality

PLS was the method of choice for multivariate regression for many years, but OPLS has lately seen an increase in metabolomics data analysis, especially for discrimination and biomarker discovery. The reason is that although the methods explain the same variance in both **X** and **Y** matrices and have the same predictive capability,

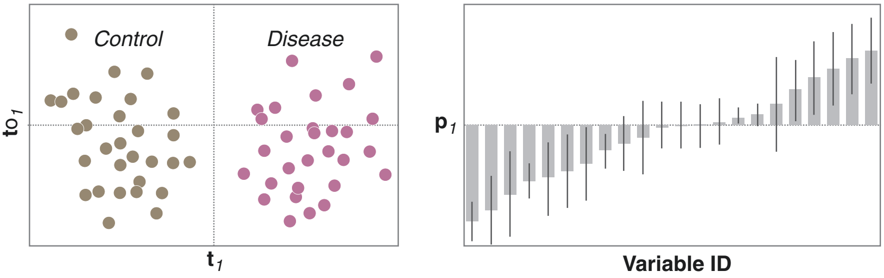

Fig. 7.3 OPLS-DA predictive vs orthogonal scores (*left*) and predictive loadings (*right*) for a two-class separation (e.g. control vs disease). The disease group, with positive predictive scores, has higher values than the control group in the variables with positive $\mathbf{p}_1$ (on the *right* of the loadings plot); it has lower values than the control group in the variables with negative $\mathbf{p}_1$ (*left side* of the loadings plot). The loadings weights were ordered according to magnitude, for easy visualization of its importance, and also the existence of confidence intervals which indicate their statistical significance

PLS computes latent variables that contain mixed sources of variation, while OPLS decomposes the structured variation into predictive and orthogonal. In the simplest case, OPLS with only one y variable – or OPLS-DA with two classes – the information related to **y** is contained in the first predictive component, while the orthogonal components contain information related to other sources of structured variance, while discarding residual variance or noise.

It is important to realize though that orthogonal components contain information that is not noise [25] and should be investigated in order to bring more understanding of an experiment. With that in mind, the datasets should be accompanied of the most complete amount of meta-information regarding unintended sources of variation such as sample preparation, experimental conditions and characterization of samples and variables as possible. In some cases, patterns and groups of samples (or variables) can be seen in the orthogonal scores (colouring them according to the meta-information may help), which can be related to that variation, e.g. sample batch, gender, age, sample dilution or other stratifications of the data. Then the orthogonal loadings should be investigated to see which variables have influence in the orthogonal score trends and groups. As all components in the model have their variation quantified, that may allow additional understanding of the relative variation in the phenomenon in study in comparison to others and, e.g. allow better tuning of experimental conditions in future experiments.

7.4.5 Cluster Analysis

Many cluster analysis methods exist, because as some authors consider, “clustering is in the eye of the beholder” [26]. Nonetheless, due to its simplicity and usefulness, hierarchical cluster analysis (HCA) has been widely used and will thus be presented. This is a non-supervised clustering method, used to put in evidence natural clustering of samples and/or variables, in the dataset. In case both samples and variables are clustered, one can see which clusters of variables are defining the clustering of the samples. Although the method is generally used for multivariable analysis, its nature is not multivariate, as no latent variables are defined.

In one of its forms, the method starts by considering that each single object is a cluster. On the first iteration, it finds the minimal distance between two of these (single object) clusters and clusters them. In the second iteration, it finds again the minimal distance between the updated clusters and clusters them. It proceeds the same way until all objects are part of the same cluster. Thus, since the beginning (after appropriate normalization/transformation of the objects in study), two parameters must be defined: the distance metrics to use and the linkage type. Distance metrics is related to how one measures "closeness" of two objects, and commonly used metrics are the Euclidean and Mahalanobis distances, or the Pearson and Spearman correlations. Linkage type is related to which objects in the current groups are used to calculate those distances, and common types are "single" (minimum distance between one object in each group), "average" (distance between averages of the objects in the groups) and "Ward" (minimum model error increase for merging two clusters). A dendrogram of the clustering process can be plotted, in which the length of the bars represent the distance between the clusters, together with a heat map of the actual data values (see Fig. 7.4).

Considering the samples, and depending on the study context and objective, the method can be applied to the actual data (metabolic features values), to its PCA scores, PCA distances to model, or any other meaningful transformation of the data.

The major advantages of the method are that it is easy to understand and its application is straightforward. The major disadvantage is the difficulty in interpreting the data when there are too many samples or too many variables (most common in metabolomics).

7.4.6 Independent Component Analysis (ICA)

ICA is a blind source separation method used in signal processing, and it separates multivariate signals into additive subcomponents. Its interpretation is similar to PCA, but instead of orthogonal components, it calculates non-Gaussian, mutually independent ones. Contrary to PCA, ICA does not order the components according to variance, and the number of components influences the structure of the components themselves; thus an adequate determination of the right number of components is of utmost importance. ICA algorithms have been used for analysis of metabolomics data, to detect metabolic patterns [27], phenotypes [28], and in class discrimination [29].

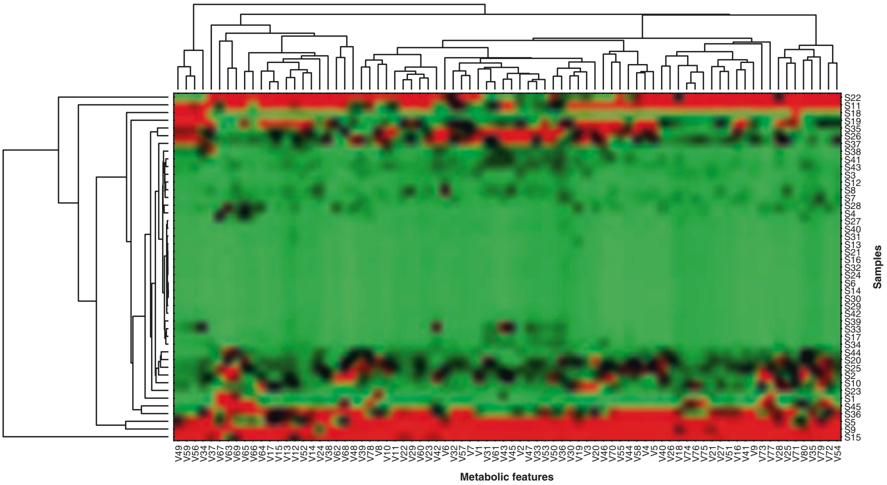

Fig. 7.4 Example of heat map and dendrograms clustering samples and metabolic features in a dataset after HCA using Euclidean distance and average linkage. By looking at the heat map, one can have visual clues about which metabolic features are aggregating the samples in clusters. In the case of the samples, the horizontal lines in their dendrogram (on the *left*) are proportional to cluster distance; for metabolic features (on the *top*), it is the vertical bars

7.5 Complementary Strategies for Data Analysis

Most applications of multivariate analysis in metabolomics use PCA for data exploration and then OPLS or OPLS-DA for regression or class discrimination/biomarker discovery, respectively. The methods are applied directly to the dataset itself, after appropriate preprocessing. Nonetheless, these same methods can be used to analyse or visualize the data in creative and helpful ways, depending on the experimental design and the objectives of the study. Below we present some of those examples.

7.5.1 OPLS-DA Strategies for Comparison of Discriminant Metabolites

As referred in Sect. 7.4.3, OPLS-DA can be used for multiple class discrimination and classification, but the direction of class separation may not align over the latent variables axes, thus turning interpretation less straightforward. In case the objective of the study is to understand the difference of multiple classes (treatments, conditions or disease states) to the same control class, it may be preferable to model each of the classes against control separately (e.g. control vs disease A and control vs disease B) and then compare the results. Two suggestions on how to do that are presented below:

(a) *Comparison of models of two classes vs same control*:

In this case, one can use the shared and unique structures (SUS) plot [30]. For the two-class case OPLS-DA models always represent the class discrimination along the predictive ("abscissa" axis) component, which allows straightforward loadings interpretation. For more than two classes/properties, this representation may not be possible to do using only one predictive component; thus the class separation may not be along a single axis. For this reason, the most convenient way of comparing two models is to create individual models for each of the possibilities (e.g. control vs disease A and control vs disease B) and then compare their loadings against each other.

From the OPLS (DA) models, different loadings vectors can be obtained. The correlation between the predictive score vector and each of the **X** variables is defined as the p_{correl} loadings. Being composed of actual correlations, its values vary between −1 and 1, thus appropriately standardized for inter-model comparison. The SUS plot is simply a scatter plot of the p_{correl} of two individual models. It should be visualized simultaneously with a p-loadings plot with confidence intervals (or any other measure of variable significance), so one can also see which variables are significant. Three different situations may arise for each of the significant variables:

(i) If aligned along a positive "/" diagonal, they show the same behaviour in both models (e.g. increased concentration of metabolite $\mathbf{X}_i$ in disease A vs control as well as in disease B vs control).

(ii) If aligned along a negative "\" diagonal, they show opposite behaviour in each model (e.g. increased metabolite concentration of metabolite $\mathbf{X}_i$ in disease A vs control, but decreased in disease B vs control).

(iii) Aligned along the horizontal/vertical axis shows an effect in one of the models, but not in the other.

(b) *Comparison of models of more than two classes vs same control*:

With more than two classes, the SUS plot approach gets complicated. A better visualization approach can be made using a network approach, plotting together the significant variables from each of the models. When considering, e.g. a small number of different diseases in relation to the same control class, the following definitions may be used (see Fig. 7.5):

(i) The different diseases are represented as central nodes, in a different shape/size and colour than the metabolites.

(ii) For each disease, its significant metabolites are represented as peripheral nodes, connected through directed edges to the disease.

(iii) The direction (or colour) of the edges indicates if the metabolite is more concentrated in the disease (pointing to the disease) or in the control (pointing to the metabolite).

(iv) The edge width can be used to denote the degree of variable significance (*p*-value, correlation, fold change).

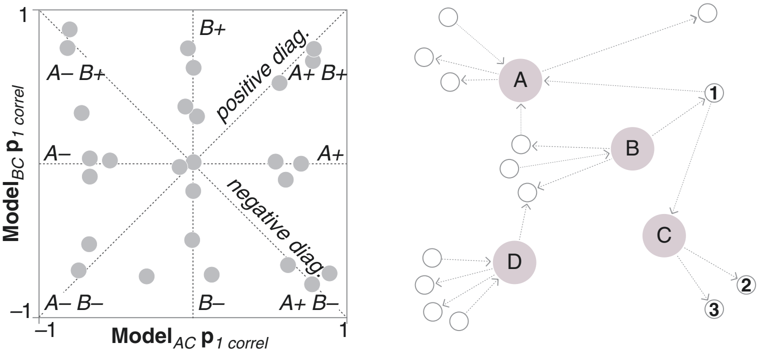

Fig. 7.5 Plots for comparing metabolites coming from models from, e.g. different diseases vs the same control group

(v) Colour codes can be used for each metabolite, reflecting the number of diseases they are common to or any other relevant information (e.g. chemical class).

As shown in Fig. 7.5 (left), SUS plot is a scatter plot of p_{correl} of two models of, e.g. diseases A and B vs control (C). Each p_{correl} is a correlation value, thus varying between −1 and 1 in each model. Variables (features or metabolites) over the positive diagonal are upregulated (A + B+) or downregulated in both diseases (A−B−) in relation to control. Variables over the negative diagonal have inverse behaviour in each of the diseases (A + B−) or (A−B+) in relation to control. Variables over the *x*-/*y*-axis are only up-/downregulated in one of the diseases (e.g. A+ or B+), but not significantly different from control in the other.

Figure 7.5 (right) shows a network representation of relevant metabolites (small circles) obtained for OPLS-DA for comparisons of, e.g. different disease classes (A–D, large circles) to the same control. The colour and size of the nodes differentiate the disease classes from the metabolites. The edge arrows pointing from a metabolite to one of the disease nodes (A–D) indicate that the metabolite level is higher in that disease than in the control group and vice versa. As examples, metabolite 1 is statistically significant in the models of diseases A, B and C. Looking at the arrow directions, it is upregulated in A and C and downregulated in B, in relation to control. Metabolites 2 and 3 are only downregulated in the model of disease C, in relation to control.

In case there are not many variables, but many classes, one can change the roles of classes with metabolites in the plots.

7.5.2 Comparisons of Trajectories and Profiles

The modelling of multivariate metabolic trajectories has been used mostly to follow time series processes using different strategies or in different subjects. It can be applied when the design of experiments uses groups of samples that have a sequential dependence (e.g. follow multiple individuals during time). The main idea is to use multivariate modelling to follow the evolution of each of the groups of dependent samples and then compare their score trajectories. Although not all these examples are in the clinical context, they may be adapted to it, if the right experimental design is used.

Using a method called "scaled-to-maximum, aligned and reduced trajectories" (SMART) [31], two groups of animals were studied in relation to the effects of drugs against control using ^{1}H NMR. The average of the initial time point of each individual is subtracted from all the samples of all individuals, to achieve a similar start point. The data is then scaled to a common magnitude by using the largest magnitude value for each treatment group, prior to using PCA (scores) to visualize the average trajectories for each treatment. Similar trajectories correspond to similar behaviour of the groups and vice versa. The same strategy can be used in other experimental settings.

In another type of application, urine samples of patients following a kidney transplant were collected in time and studied using ^{1}H NMR, in order to identify profiles for toxicity/rejection or normal recovery [32]. Grafts take different time in different individuals until actually start working properly/incorrectly. In this type of analysis, each patient was used as their own reference; thus the specific changes occurring during time for each individual could be examined. As the samples of each individual were not separated into "before" and "after" graft classes, the first objective was to identify time samples related to those two moments in the procedure. For that the researchers first used PCA in each patient and selected grouped samples in the extreme sides of the scores into the two classes ("before" and "after" graft functioning). Then they performed OPLS-DA to discriminate between those classes for each patient and collected the predictive discriminant loadings into a matrix. Finally, these discriminant loadings – which represent the profile that discriminates "before" and "after" graft for each patient – were used to represent each of the patients. PCA was performed in this matrix of loadings profiles to find a common effect profile and the most important metabolites for good patient recovery or kidney rejection.

Finally, we discuss a high-throughput multiple comparison study of transgenic tree lines against a common wild type done over several years [33]. The study objective was to understand which mutant lines (around five biological replicates per line) were more affected by the genetic engineering and which ones were similar in metabolic features. A batch effect due to biological and experimental variation did not allow the global comparison of phenotypes using the original data. The authors used an integrated chemometrics pipeline for the analysis of the data, which had the additional advantage of reducing those batch effects. This pipeline started with PCA for quality control of the wild-type samples of each of the batches separately, to detect outliers. Then, they investigated the existence of outliers in the mutants, by projecting the mutant plant samples of each batch in their respective batch PCA and looking at how they differed from the ones in their group. After the data was cleaned of outliers, they used OPLS-DA for class discrimination between the control samples and each of the mutant lines. This had the objective of finding the pattern (OPLS-DA predictive loadings) for differentiation from control for each of the mutants and also of reducing the batch differences for global comparison. The loadings representing the differentiation pattern were used in representation of each mutant line and visualized using PCA and HCA, where clusters of mutants could be visualized, together with correlated groups of metabolic features.

7.5.3 Modelling Designed Data (ASCA)

The well-established analysis of variance (ANOVA) is an ensemble of univariate methods used to analyse differences among group means. It partitions the variation of designed factor treatments and interactions, to evaluate if any of the levels in a factor or interaction is statistically different from the others. ANOVA-simultaneous component analysis (ASCA) [34, 35], in which ANOVA and PCA work together, is designed with similar intent, generalized for the multivariate case. In experiments where there is the possibility to design a balanced experiment, it can be used to evaluate which factors and interactions are statistically significant and to find which metabolites are relevant in each of those factors. Different types of data scaling may be used to amplify some aspects of the data, thus giving rise to different solutions [36].

An example of this type of design would be an experiment in which different drug formulations are given to individuals and their metabolic profile is evaluated at several time points, in order to understand if a formulation level at a specific time is statistically different from the others.

ASCA tests for the statistical significance of levels' difference in multivariate factors using a random permutation approach [37] for the factor(s) of interest. The rationale is that if no level is different from the others, the averaging process will make the factors approach zero, which should happen in the permuted models, but not in a statistically significant factor. For each factor testing, the sample group is randomly changed a large number of times and each time the factors are recalculated. A *p*-value for the significance of each factor can be calculated, based on the frequency of (number of times) factors that have a sum of squares (SSQ) larger than the original, non-permuted, factor.

7.5.4 Evaluate Effects on Matched Samples

In case the objective is to study the effect of a treatment, and there are matched samples of the type "before" and "after", the strategy OPLS-effect projections (OPLS-EP) [38, 39] can be used.

The method can be seen as a generalization of t-test for paired samples, thus similar to investigating if an average if different from zero (in opposition to t-test for unpaired samples, in which the difference between averages of two classes is evaluated). It simply uses subtraction of the "before" sample from the respective "after" sample and considers that this difference will reflect the effect of the treatment. If MS instruments are used, it has also additional advantages on the reduction of batch and drift effects, attained through running the paired samples close to each other in the run order, while randomizing its relative position. The assumption is that if paired samples are ran close to each other, there is no significant drift between them.

In this strategy, the resulting "after-minus-before" subtracted data is modelled using OPLS, using (in general) metabolite data divided by its standard deviation, against a *y* vector with 1 in all its positions. While OPLS-DA on the same data would model class discrimination (comparable to unpaired t-test), OPLS-EP models effect difference (comparable to paired t-test). By plotting the predicted effect ($\mathbf{Y}_{\mathrm{hat}}$, target value of 1) for each sample, one can understand which samples had larger ($\mathbf{Y}_{\mathrm{hat}} > 1$) or smaller ($\mathbf{Y}_{\mathrm{hat}} < 1$) effect, while looking at the predictive loadings indicates which variables were more important in that effect. The advantage of the method is that it looks for an effect, without being in reality a supervised method, as the samples have no need for class definition.

7.6 OPLS-Type Model Validation

PCA is mostly used as an exploratory method, and as the inclusion of more components has no influence in the previous ones, most times there is no need to decide the appropriate number of latent variables to use in the model. The same is not true for OPLS-type models. Furthermore, once an OPLS-type model is established, before it can be used for prediction, or to decide on the significance of the discriminant variables, rigorous validation must be performed [40]. It is worth to mention that many times the score plot will look like indicating a good class separation, but later validation does not confirm that.

The methods described in the following sections are used in the context of multivariate analysis, mostly of the OPLS-type, for the selection of an appropriate number of latent variables and for model validation.

7.6.1 Internal Cross Validation (CV)

During model building, CV is generally used in order to decide on the appropriate number of latent variables to include in the model. For each number of latent variables desired, *X* is divided into subsets of samples and then one model is built at a time, containing all samples except the ones in the corresponding subset. The subset samples are then predicted in the corresponding model, and the difference between the expected and the predicted value is saved. By doing the same for all subsets, one can obtain the predicted residual error sum of squares (PRESS) and additional statistics that allow model quality comparison. The number of latent variables that gives the least prediction error is selected for the final model. A root mean squared error of cross validation (RMSECV) can be calculated and expressed in the same units as the **Y** variable. One should then evaluate three important statistics:

(i) R2X: fraction of variance of **X** explained by each latent variable. Always increases with increasing number of components, even when overfitting by modelling noise in **X**. Answers questions (in OPLS-DA) of the type "how much of the variation in **X** is related to the difference between the classes?"

(ii) R2Y: fraction of variance of **y/Y** explained by each latent variable. It always increases with increasing number of components, as it starts modelling noise in **X** in order to explain **y/Y**. This statistics answers questions (in OPLS-DA) of the type "how good is the separation between the two classes?"

(iii) Q2: The most important statistic to decide on the quality of the model, it varies between[-INF, 1]. It is the fraction of variance of **y/Y** predicted by each latent variable. Because it is based on prediction, the inclusion of noise should not increase Q2. However, although it is expected to stop increasing, or to start decreasing, after all structured information was modelled, that is not always observed. This statistic answers questions (in OPLS-DA) of the type "how well can we predict the two classes?"

CV yields different statistic results, depending on how CV groups are defined. A commonly used strategy of leaving one sample out (LOO) at each CV round is not advised [41], as the perturbation in the data may be too weak to have a significant effect. CV should also not be used with replicate samples, as the inclusion of one of the replicates allows better predictions of its sisters, and this will inflate the statistics, showing better results than it should. Designed data may also have its problems, as the removal of some influential samples from a model may destroy the structure of the data and not allow them to be well predicted (e.g. as for samples in the extremes of the design factors).

In general, CV rounds should be defined in a balanced way, with each round containing samples from all quadrants of the experimental design. When dealing with datasets containing multiple individuals and samples of, e.g. different disease phases of the same individual, a practical advice would be to leave all samples from one individual out in each of the CV rounds, due to risk of autocorrelation. This strategy removes the contribution of each individual to build a model in each round of CV, while evaluating how the samples of that individual match the other individuals for each of its disease phases.

7.6.2 Cross Validation Scores (CV-Scores)

During CV, the different samples' subsets are predicted, and scores can be obtained. Because they are predicted from samples that were not used in the same round to build an OPLS-DA model, classes in the CV-scores always look less separated than when visualizing the scores (obtained from the model including all samples). In the case of an OPLS regression model, the CV-scores will look less correlated to **y**. Nevertheless, CV-scores should always be visualized as they give a more realistic figure of future model prediction quality. Evaluation of CV-scores is in most literature just visual, but numerical measures can also be adopted [42].

7.6.3 Cross Validation-ANOVA (CV-ANOVA)

Analysis of variance can be used to compare the size of the residuals of two models applied to the same data [43]. Here it is adapted as a diagnostic method to evaluate the reliability of an OPLS-type model. It compares the *y* predicted residuals of the model with the variation around the global average, using an F-test for comparison of variances. In case the model predicted residuals are significantly smaller than the variation around the average, the null hypothesis of equal residuals of the two models is rejected, with a certain confidence level (e.g. 0.05). It is a rapid method as it uses values calculated during cross validation, and easy to evaluate, as it provides a significance *p*-value. Nevertheless, according to the author's experience, due to biases and unidentified structured information in the data, if CV-ANOVA indicates a bad model, that is most certainly true, while if it indicates a good model, that may not necessarily be the case.

7.6.4 Permutation Test

Permutation test is a method to evaluate the statistical significance of the estimated predicted power, Q2 [22, 44, 45]. In this method, the R2Y (self-prediction) and Q2 (cross-validated prediction) of a defined model are compared to the ones from a large number of models in which the *y* vector has been randomized, and no good prediction capability is expected. The evaluation of the model validity is done by looking at the number of "random" models that present better statistics than the one being evaluated, or by looking at the intercept of the linear regression of each of those two statistics [22].

Notice that the R2Y and Q2 values are plotted in function of the correlation of each of the randomized *y* vectors and the actual true *y* vector. In case high Q2 values are found for some "random *y*" models, one should examine if that correlation is high, in which case means that the randomization process created a "randomized" y that is very similar to the actual *y* vector.

7.6.5 External Validation

While the above validation methods can provide an idea of the quality of the models, prediction of an external data can elevate our confidence in the model quality to a higher level. Depending on the objective of the experiment, more confidence can be deposited in a model that can predict samples that were acquired or processed in different times, machines, by different operators, etc. and were not used for model construction. Sensitivity and specificity can be evaluated in case of OPLS discriminant analysis/classification, while continuous measures of prediction error can be calculated in case of OPLS regression (root mean squared error of prediction, RMSEP).

7.6.6 Comparison of Model Loadings

It is itself an external validation; one can evaluate the validity of models by confirming the statistical significance of the discriminant variables. If an experiment is repeated, the class discrimination should be influenced by the same variables. An SUS plot of the two models may be used for that, in which case the same discriminant variables should be aligned along a positive diagonal.

7.6.7 Receiver Operating Characteristic (ROC) Curves

A well-established technique in clinical essays [46], it is very useful to compare over different classification models (using the area under the curve, AUC) or to evaluate thresholds for better sensitivity or specificity (using graphical representation). In the case of two-class OPLS-DA, after a model is calculated and the predicted classes obtained, a ROC curve can be calculated by incrementally moving the discrimination threshold between 0 and 1 and plotting the results for each incremental value.

7.7 Significance of Variables in OPLS-Type Regression/ Discrimination

Once an OPLS-type model has been determined and adequately validated, it is in general of interest to find out which variables (features or metabolites) are more influential in the model and to decide on its statistical significance. In order to visualize the influence of a variable in the model (regression or class discrimination), one can just sort the relevant vectors (p-loadings or VIP) by magnitude. To determine the validity of each of the variables, several strategies are described in the following sections.

7.7.1 p-loadings with Confidence Intervals

The p-loadings of the predictive latent variable represent the influence of the variable in the OPLS regression/discrimination. Furthermore, during internal cross validation (CV), multiple OPLS-type models are produced, while leaving some samples out in these different CV rounds. The p-loadings obtained in each of these CV rounds can be averaged and a standard error (error bars, confidence intervals) calculated with a predefined level of significance (e.g. 0.05). Some authors sort the p-loadings by magnitude just for model-influence visualization. Then, for statistical significance consider that, for a certain variable, if the error bars do not cross zero, the variable is statistically significant (in other words, the absolute value of variable minus standard error is larger than zero).

7.7.2 p_{correl} and Correlation Threshold

The *p*-loadings can be rescaled as the correlation coefficient between the variables in *X* and the scores (*t*) of the OPLS-type model (here defined as $\mathbf{p}_{\mathrm{correl}}$). These $\mathbf{p}_{\mathrm{correl}}$ are correlation values and, thus, have values between the limits [−1–1]. A correlation threshold for a desired level of significance (e.g. 0.05), dependent on the number of samples, can be obtained from a table of critical values for Pearson correlation. The statistically significant variables are the ones which absolute $\mathbf{p}_{\mathrm{correl}}$ larger than the adequate critical correlation threshold.

7.7.3 Variable Importance on the Projection (VIP)

This is an established and compact parameter used to summarize the importance of each of the **X** variables in a PLS with >2 components. Important variables have VIP larger than 1, while a variable is more irrelevant the lower than 1 is its VIP. There are different VIP measures for OPLS [47, 48], and researchers adopt in general the one defined in their software package.

7.7.4 Note on Significance of Variables

Some authors choose to select statistically valid features or metabolites only if they obey multiple criteria, including some of the ones described above plus others coming from univariate testing. These can be a minimum correlation needed, *p*-values after some multiple testing correction or fold change.

7.8 Univariate Analysis

Univariate analysis has been used in conjunction with multivariate analysis to study variation and to test statistical significance of parameters and variables in metabolomics studies. Notice that while multivariate analysis can handle certain amounts of batch and drift variation, univariate analysis should only be used if correction for these effects is satisfactory. Nevertheless, new attention has been given to the analysis of metabolomics data using univariate analysis [49], especially in the field of epidemiology [50]. Until recently, rare – if any – metabolomics studies were composed of thousands of samples due to its cost but also to issues related to process automatization, data handling, processing and reproducibility, among others. However, some large-scale metabolite profiling studies have now been done [51], mostly in epidemiological research, as metabolomics is expected to measure environmental and exogenous exposures more precisely than traditional questionnaires. In these studies, linear or generalized linear models are used in univariate analysis fashion, while correcting for confounders. These confounders are experimental factors that may be correlated with the outcome and in that case are not removed using OPLS-type multivariate methods.

A word should be said in relation to the use of parametric (e.g. t-tests, Pearson correlation) or nonparametric (e.g. Mann–Whitney U test, Spearman correlation) strategies. For normal populations, parametric tests are more powerful than non-parametric ones, but that is not the case for non-normal populations, unequal variances and unequal small sample sizes, where using a nonparametric test would be advantageous. While testing for normality distribution in four datasets, some authors found in average 65 % of metabolic features met normality and equality of variance assumptions. Still, as it was dependent on the dataset, they suggest to use both strategies, and if there is a large difference in the results, one should look for outliers in the dataset [49].

7.9 Multiple Testing Corrections

To decide on the statistical significance of a feature or a metabolite, e.g. if it is or not correlated with an outcome or if it has discriminant capacity between two classes, univariate methods rely on p-values. Because in metabolomics untargeted studies one is looking after thousands of variables, multiple testing corrections must be applied. The Bonferroni correction was commonly used in the past, when the number of variables was small, but as it corrects for the family-wise error rate (FWER) – the probability of at least one false positive – it ends up being too conservative. Benjamini–Hochberg and other corrections that control the false discovery rate (FDR) [52, 53] are less conservative and widely applied. Nonetheless, due to the high degree of correlation observed in metabolomics data (e.g. adjacent intensities in NMR peaks), existence of multiple features for the same metabolite (e.g. LC-MS dimers, adducts; NMR signal multiplicity), they are also considered not appropriate. Thus, permutation strategies such as the metabolome-wide significance level (MWSL) [54] have been developed to control for the FWER, which determine more robust p-value thresholds for discovery than the above methods.

7.10 Practical Aspects of Chemometrics in the Context of Preprocessing, Pretreatment and Experimental Design

Many decisions must be taken when designing a metabolomics experiment, regarding sample type and number, cost, time, human resources, instruments and data analysis methods.

The decisions taken will provide answers to different questions; thus a very well-defined idea about the methods that will be used for data analysis and interpretation is fundamental, in order to be able to pose objective questions and obtain correspondingly appropriate answers. This has a retrospective impact on the design of experiments itself. Although not the main focus of this chapter, it seems appropriate to include a brief description of different options that can be made, which condition the chemometrics data analysis and may be used for batch correction in MS data.

7.10.1 ^{1}H NMR Data: Types of Data Matrices

Different strategies can be used for the preprocessing of ^{1}H NMR data, depending on the type of sample (e.g. urine or serum), because of chemical shifts due to physicochemical sample differences. Blood samples are expected not to change much due to homeostatic regulation, while urine is known to change more in concentration as well as in properties, as it is more affected by microbiota, drugs, diet and disease [50, 55].

Direct analysis after alignment: if the alignment is good enough, the data can be immediately analysed. When using multivariate methods for the data analysis, data tend to be scaled by mean centring or Pareto normalization, with optional log transformation, so the spectral structure is kept. In this case, unit variance normalization is not used as it gives the same importance to variation in the signal and in the noise region, and the loadings do not show spectral structure similar to the data. If large "saw-tooth" (inverted peaks) effects are found in the loadings, it is a sign that the alignment was not performed perfectly.

Analysis after alignment and binning: if there are some issues with the alignment, but the whole spectrum is to be analysed, binning adjacent values can be used. If multivariate analysis is used, data should be normalized by centring or Pareto normalization, to keep the original spectral structure, with optional log transformation.

Analysis after alignment and peak picking: a similar strategy to binning is to integrate peaks, but in a more targeted way, thus rejecting noise regions. If multivariate analysis is to be used, unit variance scaling can be used, as noise regions are not supposed to exist. When using this strategy, the spectral structure is lost.

7.10.2 MS Data: Reducing Batch and Drift Effects

GC/LC-MS instruments are prone to batch and time drift effects, due to changes in instrument sensitivity and intensity, among other effects. Targeted methods correct batch and drift effects with the inclusion of labelled internal standards, with which a ratio between the target compound and the internal standard can be calculated. For untargeted GC-MS and LC-MS, several strategies have been used to correct for these effects: inclusion of periodic quality control samples (pooled from all or from a group of samples in the experimental set) that are expected to yield the same results along time, addition of internal standards to the samples (heavily labelled and/or not occurring in the samples) representing different compound classes, and experimental design using paired samples (when having samples with and without effect). These methods present both advantages and disadvantages, which are tentatively explained below.

Periodic Internal Quality Control (iQC) Samples [56] An adequate volume of pooled sample is built by pooling an amount of each of the biological samples. Subsamples of this main sample, assumedly with equal composition and concentration, are interspersed (e.g. every fifth sample) with the biological samples and ran in the instrument. For each individual variable, these iQC samples are then modelled using locally weighted regression (e.g. robust loess). Once a model is established, one can calculate the ratio between each biological sample and the LOESS curve (while the iQC samples should all equal 1) for each variable. This method is used both for batch and drift correction. While potentially the most adequate correction method, its major disadvantage is the increase in the total number of samples, with impact on time and cost of the experiment.

Internal Standards (IS) A certain number of quality control labelled compounds, not expected to have endogenous expression in the samples, representing different chemical/biological classes (e.g. amino acids, fatty acids) are added in the same concentration to each biological sample. These compounds are assumed to be in the same concentration in each sample. The advantage of the method is that there is no need for additional samples. The disadvantage is that the internal standards may not be the adequate ones to normalize the data. To correct for batch/drift using internal standards, the following strategies have been applied:

(i) *Correction by single or multiple internal standards* (*IS*): if a single IS was used, the ratio or log2 ratio between each variable and the IS can be calculated, and the variable is considered corrected. If multiple IS were used, the decision of which IS shall be used to normalize a certain variable can be done according to maximum IS correlation [57] or by minimal retention time difference [58].

(ii) *Correction using multiple IS and PCA*: the features representing the IS are normalized dividing each one by their corresponding standard deviation. PCA is calculated on this dataset, and the scores of the first component are obtained, representing the major batch/drift effect on the dataset. Then the procedure is the same as in the previous case. For each variable, the ratio between each sample and the corresponding score value is calculated, normalizing the data. The major disadvantage of this method is that it can only correct one major batch/drift effect and may even wrongly correct features that were less affected by those effects.

(iii) *Correction using OPLS*: a procedure similar in concept to orthogonal signal correction has been applied to microchannel microarray data using OPLS [59], cleaning the data from orthogonal variation that is not common within sets of biological replicates. The method uses the data in matrix *X* and a dummy matrix identifying the replicates in a matrix *Y*. After it identifies the orthogonal information, it builds the corrected data matrix using matrix multiplication of predictive scores and loadings, plus the residual variance. The strategy seems applicable to metabolomics, and a variation of it, using information from the IS samples, has been used in the context of batch normalization and drift correction [60].

Experimental Design Using Paired Samples In case there are reference samples, like "before" (baseline) and "after" treatment for the same individual, or when using multiple time points, or matched case–control, and the objective is to study an effect (e.g. of a drug, or disease). In these cases the matched samples can be ran close to each other, and the drift between them is assumed as negligible. The baseline sample can then be subtracted from the effect(s) sample(s), with the result being the difference between the two (the effect itself). Local randomization of the matched samples is used, to minimize for any sequential bias. While not needing additional samples, the major disadvantage of this method is that it is only applicable in situations where a reference sample exists. Additionally, one will be studying not the current metabolite relative levels but the effect's metabolite relative levels in relation to baseline. The OPLS-EP method previously mentioned [38] is an example of this strategy in practice, for paired samples "before" and "after" effect. Alternatively to subtracting a "before" sample (baseline), the average per group of paired samples could be subtracted from each of the respective paired samples, in which case the baseline sample would be kept for analysis.

7.11 Internet Resources and Software

Following the developments in other omics fields, efforts have been put into creating internet platforms for automated and semi-automated metabolomics data analysis. Some very good resources are now available, which may require a minimum of knowledge of the methods on the side of the researcher to output meaningful data analysis results. Among the most well known and commonly used are MetaboAnalyst [61–65], metaP-server [66], Workflow4metabolomics [67] and Galaxy-M [68], and work is in progress in the large-scale computing for medical metabolomics website PhenoMeNal [69]. Finally, the website OMICtools [70] provides a comprehensive description of software that can be used for metabolomics data analysis, as well as a number of sites that can be used for different purposes in the omics fields.

7.12 Concluding Remarks

Chemometrics has been heavily used in all steps of metabolomics studies and has here been discussed in the context of data analysis in clinical metabolomics contexts. Its relevance in this field is due to the complexity and number of variables in metabolomics datasets and the simplicity of interpretation of its results. While other strategies in bioinformatics start appearing that gather information from databases, thus needing previous identification of metabolites, chemometrics methods are purely numerical, thus finding its own place in the data analysis pipeline. The possibility of automation has brought to light some websites that provide statistical calculations, chemometrics methods included, without major input from the analyst.

Acknowledgments The author thanks Jenny Hällqvist for assistance with figures and Ibrahim Karaman for valuable discussions.

References

1. Nicholson JK, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica. 1999;29(11):1181–9.
2. Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L. Metabolite profiling for plant functional genomics. Nat Biotechnol. 2000;18(11):1157–61.
3. Massart DL, Deming SN, Michotte Y, Kaufman L, Vandeginste BGM. Chemometrics: a textbook. New York: Elsevier Sciences Ltd.; 1988.
4. Brereton RG. A short history of chemometrics: a personal view. J Chemom. 2014;28(10):749–60.
5. Piantadosi S. Clinical trials: a methodologic perspective, second edition. 2nd ed. New Jersey: John Wiley & Sons; 2005. p. 720.
6. Trygg J, Holmes E, Lundstedt T. Chemometrics in metabonomics. J Proteome Res. 2007;6(2):469–79.
7. Madsen R, Lundstedt T, Trygg J. Chemometrics in metabolomics – a review in human disease diagnosis. Anal Chim Acta. 2010;659(1–2):23–33.
8. Bartel J, Krumsiek J, Theis FJ. Statistical methods for the analysis of high-throughput metabolomics data. Comput Struct Biotechnol J. 2013;4:e201301009.
9. Pearson K. On lines and planes of closest fit to systems of points in space. Philos Mag. 1901;2(series 6, 11):559–72.
10. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemom Intell Lab. 1987;2(1–3):37–52.
11. Jackson JE. A user's guide to principal components. New York: John Wiley & sons; 1991.
12. Jolliffe IT. Principal component analysis. 2nd ed. New York: Springer-Verlag New York, Inc.; 2002.
13. Pinto J, Barros AS, Domingues MR, Goodfellow BJ, Galhano E, Pita C, et al. Following healthy pregnancy by NMR metabolomics of plasma and correlation to urine. J Proteome Res. 2015;14(2):1263–74.
14. Hotelling H. The generalization of student's ratio. Ann Math Stat. 1931;2(3):360–78.
15. Berkane M, Bentler PM. Estimation of contamination parameters and identification of outliers in multivariate data. Sociol Methods Res. 1988;17(1):55–64.

16. Filzmoser P, Ruiz-Gazen A, Thomas-Agnan C. Identification of local multivariate outliers. Stat Pap. 2014;55(1):29–47.
17. Magis D, De Boeck P. Identification of differential item functioning in multiple-group settings: a multivariate outlier detection approach. Multivar Behav Res. 2011;46(5):733–55.
18. Rocke DM, Woodruff DL. Identification of outliers in multivariate data. J Am Stat Assoc. 1996;91(435):1047–61.
19. Hubert M, Rousseeuw PJ, Vanden BK. ROBPCA: a new approach to robust principal component analysis. Technometrics. 2005;47(1):64–79.
20. Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemom Intell Lab. 2001;58(2):109–30.
21. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). J Chemom. 2002;16(3):119–28.
22. Eriksson L, Byrne T, Johansson E, Trygg J, Wikstrom C. Multi- and megavariate data analysis basic principles and applications, third revised edition. Malmo: MKS Umetrics AB; 2013.
23. Bylesjo M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. J Chemom. 2006;20(8–10):341–51.
24. Yap IK, Brown IJ, Chan Q, Wijeyesekera A, Garcia-Perez I, Bictash M, et al. Metabolome-wide association study identifies multiple biomarkers that discriminate north and south Chinese populations at differing risks of cardiovascular disease: INTERMAP study. J Proteome Res. 2010;9(12):6647–54.
25. Pinto RC, Trygg J, Gottfries J. Advantages of orthogonal inspection in chemometrics. J Chemom. 2012;26(6):231–5.
26. Estivill-Castro V. Why so many clustering algorithms: a position paper. ACM SIGKDD Explorations Newsletter. 2002;4(1):65–75.
27. Li X, Hansen J, Zhao XJ, Lu X, Weigert C, Haring HU, et al. Independent component analysis in non-hypothesis driven metabolomics: Improvement of pattern discovery and simplification of biological data interpretation demonstrated with plasma samples of exercising humans. J Chromatogr B. 2012;910:156–62.
28. Liu Y, Smirnov K, Lucio M, Gougeon RD, Alexandre H, Schmitt-Kopplin P. MetICA: independent component analysis for high-resolution mass-spectrometry based non-targeted metabolomics. BMC Bioinf. 2016;17:114.
29. Monakhova YB, Godelmann R, Kuballa T, Mushtakova SP, Rutledge DN. Independent components analysis to increase efficiency of discriminant analysis methods (FDA and LDA): Application to NMR fingerprinting of wine. Talanta. 2015;141:60–5.
30. Wiklund S, Johansson E, Sjostrom L, Mellerowicz EJ, Edlund U, Shockcor JP, et al. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. Anal Chem. 2008;80(1): 115–22.
31. Keun HC, Ebbels TM, Bollard ME, Beckonert O, Antti H, Holmes E, et al. Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. Chem Res Toxicol. 2004;17(5):579–87.
32. Stenlund H, Madsen R, Vivi A, Calderisi M, Lundstedt T, Tassini M, et al. Monitoring kidney-transplant patients using metabolomics and dynamic modeling. Chemom Intell Lab. 2009;98(1):45–50.
33. Pinto RC, Gerber L, Eliasson M, Sundberg B, Trygg J. Strategy for minimizing between-study variation of large-scale phenotypic experiments using multivariate analysis. Anal Chem. 2012;84(20):8675–81.
34. Smilde AK, Jansen JJ, Hoefsloot HC, Lamers RJ, van der Greef J, Timmerman ME. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. Bioinformatics. 2005;21(13):3043–8.
35. Jansen JJ, Hoefsloot HCJ, van der Greef J, Timmerman ME, Westerhuis JA, Smilde AK. ASCA: analysis of multivariate data obtained from an experimental design. J Chemom. 2005;19(9):469–81.

36. Timmerman ME, Hoefsloot HC, Smilde AK, Ceulemans E. Scaling in ANOVA-simultaneous component analysis. Metabolomics. 2015;11(5):1265–76.
37. Vis DJ, Westerhuis JA, Smilde AK, van der Greef J. Statistical validation of megavariate effects in ASCA. BMC Bioinf. 2007;8:322.
38. Jonsson P, Wuolikainen A, Thysell E, Chorell E, Stattin P, Wikstrom P, et al. Constrained randomization and multivariate effect projections improve information extraction and biomarker pattern discovery in metabolomics studies involving dependent samples. Metabolomics. 2015;11(6):1667–78.
39. Bjorkblom B, Wibom C, Jonsson P, Moren L, Andersson U, Johannesen TB, et al. Metabolomic screening of pre-diagnostic serum samples identifies association between alpha- and gamma-tocopherols and glioblastoma risk. Oncotarget. 2016; 7(24):37043–37053.
40. Szymanska E, Saccenti E, Smilde AK, Westerhuis JA. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. Metabolomics. 2012;8 Suppl 1:3–16.
41. Shao J. Linear-model selection by cross-validation. J Am Stat Assoc. 1993;88(422):486–94.
42. Worley B, Halouska S, Powers R. Utilities for quantifying separation in PCA/PLS-DA scores plots. Anal Biochem. 2013;433(2):102–4.
43. Eriksson L, Trygg J, Wold S. CV-ANOVA for significance testing of PLS and OPLS (R) models. J Chemom. 2008;22(11–12):594–600.
44. Van der Voet H. Comparing the predictive accuracy of models using a simple randomization test. Chemom Intell Lab. 1994;25:313–23.
45. Eigenvector Research I. PLS toolbox: Permutation Test: Eigenvector Research, Manson, WA, USA, Inc.; 2014. http://wiki.eigenvector.com/index.php?title=Tools:_Permutation_Test
46. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem. 1993;39(4):561–77.
47. Galindo-Prieto B, Eriksson L, Trygg J. Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). J Chemom. 2014;28(8):623–32.
48. Galindo-Prieto B, Eriksson L, Trygg J. Variable influence on projection (VIP) for OPLS models and its applicability in multivariate time series analysis. Chemom Intell Lab. 2015;146:297–304.
49. Vinaixa M, Samino S, Saez I, Duran J, Guinovart JJ, Yanes O. A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. Metabolites. 2012;2(4):775–95.
50. Tzoulaki I, Ebbels TM, Valdes A, Elliott P, Ioannidis JP. Design and analysis of metabolomics studies in epidemiologic research: a primer on -omic technologies. Am J Epidemiol. 2014;180(2):129–39.
51. Dunn WB, Lin W, Broadhurst D, Begley P, Brown M, Zelena E, et al. Molecular phenotyping of a UK population: defining the human serum metabolome. Metabolomics. 2015;11:9–26.
52. Benjamini Y, Hochberg Y. Controlling the false discovery rate – a practical and powerful approach to multiple testing. J Roy Stat Soc B Methodol. 1995;57(1):289–300.
53. Benjamini Y, Cohen R. Weighted false discovery rate controlling procedures for clinical trials. Biostatistics. 2016.
54. Chadeau-Hyam M, Ebbels TMD, Brown IJ, Chan Q, Stemler J, Huang CC, et al. Metabolic profiling and the metabolome-wide association study: significance level for biomarker identification. J Proteome Res. 2010;9(9):4620–7.
55. Bictash M, Ebbels TM, Chan Q, Loo RL, Yap IKS, Brown IJ, et al. Opening up the "Black Box": metabolic phenotyping and metabolome-wide association studies in epidemiology. J Clin Epidemiol. 2010;63(9):970–9.
56. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc. 2011;6(7):1060–83.
57. Sysi-Aho M, Katajamaa M, Yetukuri L, Oresic M. Normalization method for metabolomics data using optimal selection of multiple internal standards. BMC Bioinf. 2007;8:93.
58. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, et al. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. Anal Chem. 2006;78(2):567–74.

59. Bylesjo M, Eriksson D, Sjodin A, Jansson S, Moritz T, Trygg J. Orthogonal projections to latent structures as a strategy for microarray data normalization. BMC Bioinf. 2007;8:207.
60. Mattsson A, Karrman A, Pinto R, Brunstrom B. Metabolic profiling of chicken embryos exposed to perfluorooctanoic acid (PFOA) and agonists to peroxisome proliferator-activated receptors. PLoS One. 2015;10(12):e0143780.
61. Xia J, Psychogios N, Young N, Wishart DS. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res. 2009;37(Web Server issue):W652–60.
62. Xia J, Wishart DS. Metabolomic data processing, analysis, and interpretation using MetaboAnalyst. Curr Protoc Bioinf. 2011;Chapter 14:Unit 14 0.
63. Xia J, Wishart DS. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. Nat Protoc. 2011;6(6):743–60.
64. Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS. MetaboAnalyst 2.0 – a comprehensive server for metabolomic data analysis. Nucleic Acids Res. 2012;40(Web Server issue):W127–33.
65. Xia J, Sinelnikov IV, Han B, Wishart DS. MetaboAnalyst 3.0 – making metabolomics more meaningful. Nucleic Acids Res. 2015;43(W1):W251–7.
66. Kastenmuller G, Romisch-Margl W, Wagele B, Altmaier E, Suhre K. metaP-server: a web-based metabolomics data analysis tool. J Biomed Biotechnol. 2011; Volume 2011, Article ID 839862, 7 pages.
67. Giacomoni F, Le Corguille G, Monsoor M, Landi M, Pericard P, Petera M, et al. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. Bioinformatics. 2015;31(9):1493–5.
68. Davidson RL, Weber RJ, Liu H, Sharma-Oates A, Viant MR. Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. Gigascience. 2016;5:10.
69. Kale N, Steinbeck C, Consortium P. PhenoMeNal – an e-infrastructure for analysis of metabolic phenotype data: Metabonews. 2016. Available from: http://www.metabonews.ca/Jan2016/MetaboNews_Jan2016.htm.
70. Henry VJ, Bandrowski AE, Pepin AS, Gonzalez BJ, Desfeux A. OMICtools: an informative directory for multi-omic data analysis. Database (2014) 2014: bau069 doi:10.1093/database/bau069.