Journal of reviews proteome research

Chemometrics in Metabonomics

Johan Trygg,[†] Elaine Holmes,[‡] and Torbjörn Lundstedt*[,§,||]

Research group for Chemometrics, Institute of Chemistry, Umeå University, Sweden, Biological Chemistry, Biomedicinal Sciences Division, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, London SW7 2AZ, United Kingdom, Department of Pharmaceutical Chemistry, Uppsala University, Sweden, and AcurePharma, Uppsala, Sweden

Received November 11, 2006

We provide an overview of how the underlying philosophy of chemometrics is integrated throughout metabonomic studies. Four steps are demonstrated: (1) definition of the aim, (2) selection of objects, (3) sample preparation and characterization, and (4) evaluation of the collected data. This includes the tools applied for linear modeling, for example, Statistical Experimental Design (SED), Principal Component Analysis (PCA), Partial least-squares (PLS), Orthogonal-PLS (OPLS), and dynamic extensions thereof. This is illustrated by examples from the literature.

Keywords: Statistical Experimental Design (SED) • PCA • OPLS • Class-specific studies • Dynamic studies • Multivariate Design

Introduction

Our intention with this review is to give an overview of papers applying a chemometrical approach throughout a metabonomic study. We will provide the readers with a number of examples on different types of problems and methods applied. In all of the cases, we will only briefly discuss two or three examples and then provide the readers with references to a number of papers dealing with similar examples in an analogue way. The philosophy presented in this review strongly supports the statement made by Robertson[1] in his metabonomics review "However, the implementation and interpretation of the technology and data it generates is not something that should be trivialised. Proper expertise in biological sciences, analytical sciences (nuclear magnetic resonance and/or mass spectrometry) and chemometrics should all be considered necessary prerequisites. If these factors are properly considered, the technology can add significant value as a tool for preclinical toxicologists." We just would like to make a minor change to the last sentence of the citation; "...If these factors are properly considered, the technology can add significant value as a tool for disease diagnosis, pharmacodynamic studies, preclinical toxicologists, and more to come."

In metabonomics, as well as in other branches of science and technology, there is a steady trend toward 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., wavelength, mass number, chemical shift). This development generates huge and complex data tables, which are hard to summarize and overview without appropriate tools. However, in biology, chemometric methodology has been largely overlooked in favor of traditional statistics. It is not until recently that the overwhelming size and complexity of the 'omics' technologies has driven biology toward the adoption of chemometric methods. That includes efficient and robust methods for modeling and analysis of complicated chemical/biological data tables that produce interpretable and reliable models capable of handling incomplete, noisy, and collinear data structures. These methods include principal component analysis[2] (PCA) and partial least-squares[3,4] (PLS). It is also important to stress that chemometrics also provides a means of collecting relevant information through statistical experimental design[5−7] (SED).

The underlying philosophy of chemometrics, in combination with the chemometrical toolbox, can efficiently be applied throughout a metabonomic study. The philosophy is needed already from the start of a study through the whole process to the biological interpretation.

The Study Design: Make Data Contain Information. The metabonomics approach is more demanding on the quality, accuracy, and richness of information in data sets. Statistical Experimental Design (SED)[5,6] 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 10−20 experiments. By adding additional experiments, one can investigate factors more thoroughly, for example, the time dependence from two to five time points. In addition, the noise level is decreased by means of

* To whom correspondence should be addressed. E-mail: torbjorn.lundstedt@acurepharma.com.
[†] Umeå University.
[‡] Imperial College London.
[§] Uppsala University.
[||] AcurePharma.

10.1021/pr060594q CCC: $37.00

Published on Web 12/22/2006

averaging, the functional space is efficiently mapped, and interactions and synergisms are seen. Antti et al.[8] have applied a statistical experimental design to investigate the effect of the dose of hydrazine and time on liver toxicity. The result from the NMR and clinical chemistry was evaluated by PLS. The PLS analysis could also reveal the correlation pattern between the different blocks as well as within blocks according to dose, time, and the interaction between time and dose.

Extract Information from Data. In metabonomic studies, the observations and samples are often characterized using modern instrumentation such as GC−MS, LC−MS, and LC−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 modeling of these complex data. Principal Component Analysis[2] (PCA) is the workhorse in chemometrics. PCA makes 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, projection-based methods represent a solid basis for metabonomic analysis. Canonical correlation,[9] correspondence analysis,[10] neural networks,[11−12] Bayesian modeling,[13] and hidden Markov models[14] represent additional modeling methods but are outside the scope of this review.

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, a more focused evaluation and analysis of the data is possible. 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 modeling, this additional knowledge constitutes an extra data table, that is, a **Y** matrix. Partial least-squares[3] (PLS) and Orthogonal-PLS[15−19] (OPLS) represent two modeling 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.

Chemometric Approach to Metabonomic Studies

1. Step 1: Define the Aim. It is important to formulate the objectives and goals of the metabonomic study. A number of questions have to be answered and/or taken into consideration in both the design of study as well as in the evaluation of the outcome. For example, What is previously known? What additional information is needed? How to reach the objectives; thta is, what experiments are needed and how to perform them? If these questions cannot be answered, there is no point to continue.

2. Step 2: Selection of Objects. The selection of the objects (e.g., samples, individuals) 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,

2.1. 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).[20] Thus, projection-based methods can be regarded as a data analysis toolbox, for indirect observation of these LVs. This class of models is conceptually very different from traditional regression models with independent predictor variables. 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 multidimensional 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 to each other see Figure 1. 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 ($\mathbf{t}_1$−$\mathbf{t}_2$) 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 ($\mathbf{p}_1$,$\mathbf{p}_2$) 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 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**.

$$\mathbf{X} = \mathrm{TP}^T + E = \mathbf{t}_1\mathbf{p}_1{}^T + \mathbf{t}_2\mathbf{p}_2{}^T + E$$

2.2. 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)[21,22] is a combination of multivariate characterization (MVC),[23−25] principal component analysis (PCA), and Statistical Experimental Design (SED) to select a diverse set of objects that represents all objects, that is, spans the variation. There is a number of different experimental designs that can be applied to span the variation in a

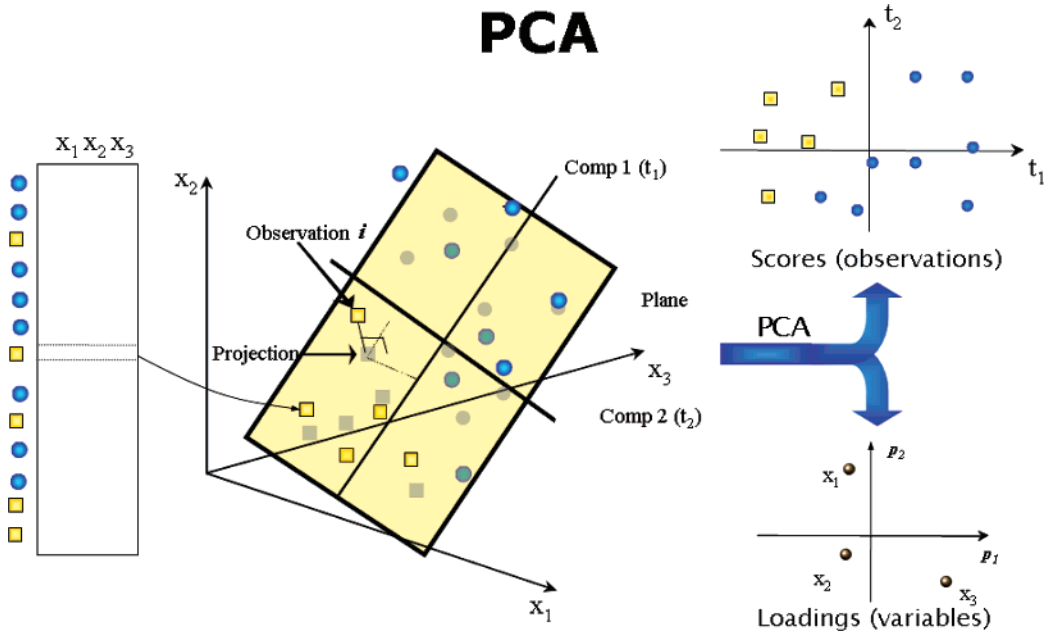

Figure 1. A principal component analysis (PCA) model approximates the variation in a data table by a low dimensional model plane. This model plane provides a score plot, where the relation among the observations or samples in the model plane is visualized, for example, if there are any groupings, trends, or outliers. The loading plot describes the influence of the variables in the model plane, 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.

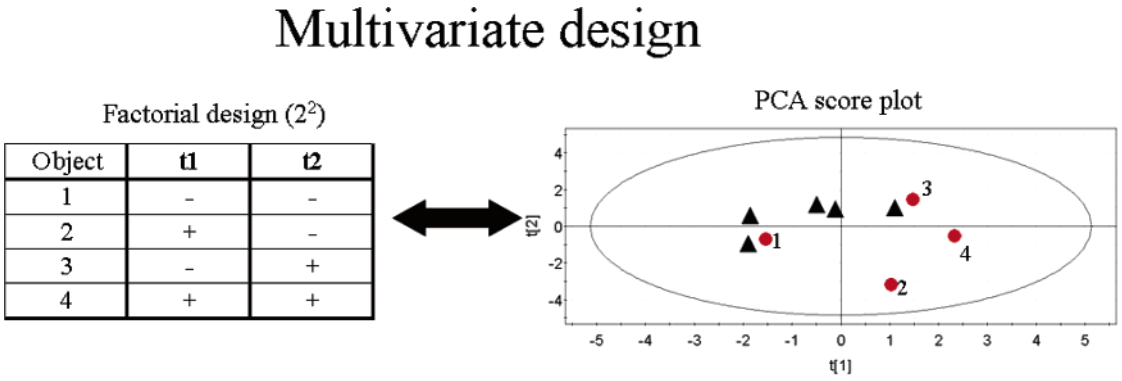

Figure 2. Four objects are selected according to a multivariate design that spans the biological variation.

systematic way and obtain well-balanced data. The most commonly used are factorial designs[6] and D-optimal design[26] that fulfill the criteria of balanced data and orthogonality. In MVD, the principal component model scores, for example, $\mathbf{t}_1$ and $\mathbf{t}_2$ are used to select the objects, see Figure 2. The selection is based on diversity between the objects.

3. Step 3: Sample Preparation and Characterization. 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 physicochemical diversity of metabolites (amino acids, fatty acids, carbohydrates, and organic acids) raises problems for extraction and working procedures for different analytical techniques. Here, statistical design of experiments represents an important strategy to systematically investigate factors and optimize the experimental protocols. Typical working procedures for NMR spectroscopy for biofluids and tissue extraction are found in Appendix 4, in the SMRS Policy document.[27] For GC−MS, see refs 28 and 29.

Dumas et al.[30]assessed the analytical reproducibility of human urine samples characterized by H-NMR in the INTERMAP study. INTERMAP was launched in 1996 to investigate the relationship of multiple dietary variables to blood pressure. The conclusion was that most errors are due to specimen handling inhomogeneity.

Multiple factor analysis (MFA) was used by Dumas et al.[31] to integrate NMR and MS data for metabolic fingerprinting on cattle treated with anabolic steroids. Only minor overlap was found in the correlation structure between the MS and NMR variables. They underline the relation of a multivariate profile not only to the biological information content, but also to its inherent signature from the analytical instrument used.

4. Step 4: Evaluation of the Collected Data. In contrast to an ^{1}H-NMR spectrum, data collected from hyphenated instruments such as GC−MS, LC−MS, and UPLC-NMR must be processed 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[32−36] that result

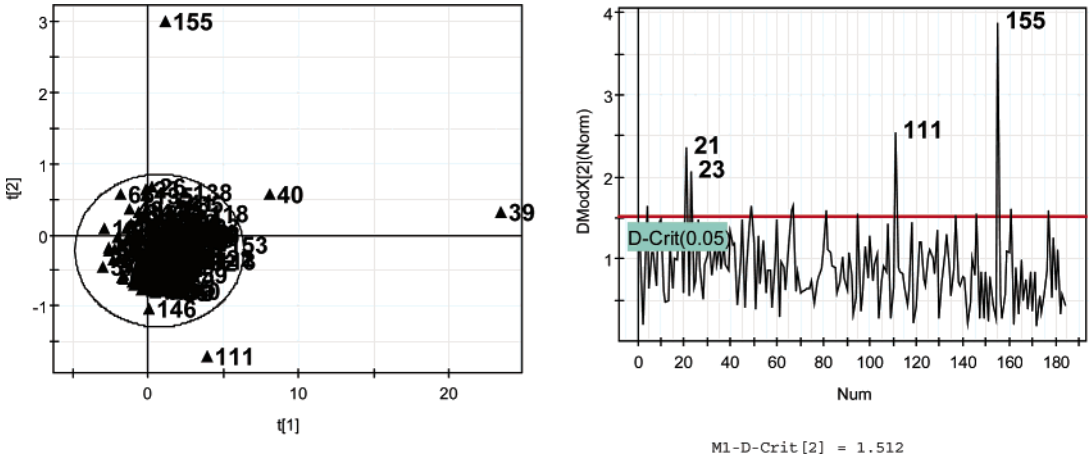

Figure 3. In the score plot (left panel), the confidence interval is defined by the Hotelling's T2 ellipse (95% confidence interval), and observations outside the confidence ellipse are considered outliers. Outliers can also be detected by the distance to model parameter, DModX, based on the model residuals (right panel).

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 cause problems for statistical modeling. Because of this, a multitude of different peak alignment methods have been developed.[37,38] Variability of chemical shifts in H-NMR spectra of biofluids, for example, due to pH variation, metal-ion concentrations, and chemical exchange phenomena, has spurred the development of bucketing and peak alignment methods. A commonly used method involves bucketing the data, where signal integration within spectral regions is performed.[39] An alternative approach has been to use automatic peak alignment methods to resolve the problem of peak position variation. Stoyanova et al.[40] removed the positional noise using PCA. Forshed et al.[41,42] and Lee et al.[43] have applied genetic algorithms to align segments of spectra to determine the misalignment across a series of NMR spectra. Cloarec et al.[18] applied the OPLS method to evaluate the influence of typical peak position variation on the robustness of pattern recognition methods and demonstrated that the inclusion of variable peak position can be beneficial and lead to useful biochemical information. 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 ($\mathbf{t}_1-\mathbf{t}_2$) 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 T^2 region, shown as an ellipse in Figure 3 (left panel), defines the 95% confidence interval of the modeled variation.[44] Outliers may also be detected in the model residuals. The distance to model plot[45] (DModX) can be used and is a statistical test for detecting outliers based on the model residual variance; see Figure 3 (right panel).

Interesting individual observations, such as outliers, can be examined and interpreted by the contribution plot.[46] It displays the weighted difference between the observation and the model center. 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.

PCA modeling was used to assess the statistical differentiation between the groups, and the covariance loadings plot for biochemical interpretation. One example is the paper by Akira et al.[47] wherein the biochemical changes between hypertensive rats and their normotensive controls to provide insight into blood pressure regulation was investigated. The design study included six male rats from each class, and urine was sampled twice at 12 and 26 weeks of age. PCA have been frequently applied in the evaluation of metabonomic data and should be the method of choice for obtaining an overview, find clusters, and to identify outliers. For a few different examples on applications see refs 48−55.

Haluska and Powers[56] have discussed the draw back of spectral noise when evaluating the data by PCA and suggest simply removing the noise regions by setting a threshold and only use the signals above the spectral noise in the PCA. Another approach to this would be to utilize the prior knowledge gained in "Study Design", which gives us the ability to separate the observations in at least two different classes, and thereby use more advanced multivariate methods such as SIMCA, PLS-DA, and/or OPLS-DA.

Class Specific Studies

Most of the published papers within the field are dealing with classification problems such as disease diagnosis or treated versus control, that is, to identify a group of control observations and another group of observations known to have a specific disease. In a number of papers, several classes can be identified, but in all of these papers, the evaluation has been made as a two-class case.

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.

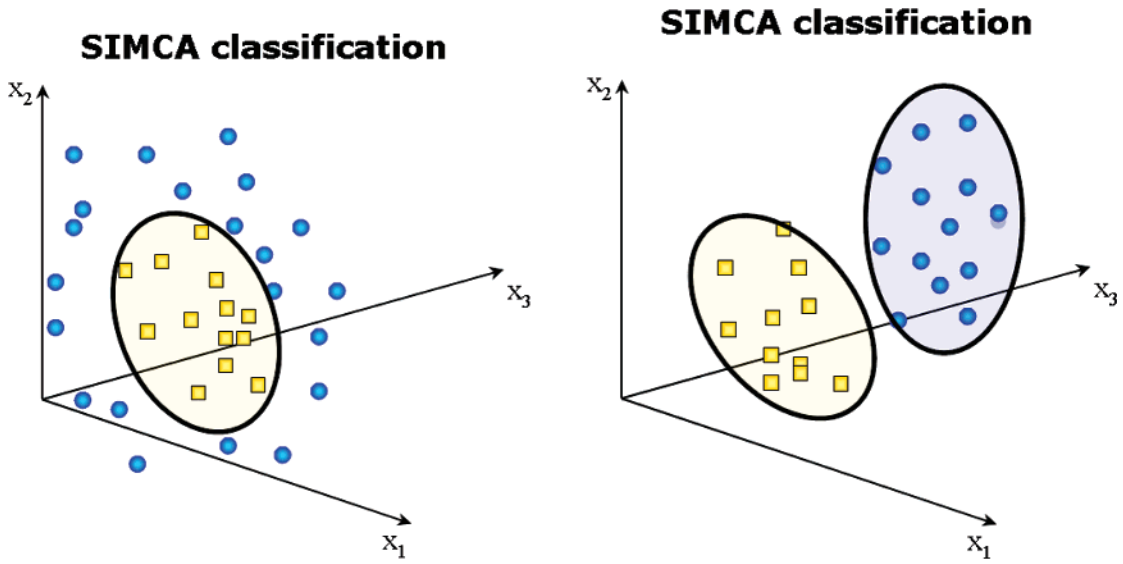

Figure 4. Illustration of SIMCA classification. In the left panel, the one-class classifier is shown, referred to as the asymmetric case. In the right panel, the SIMCA classification is shown with two classes, separately modeled by PCA.

Soft Independent Modeling of Class Analogy (SIMCA). The SIMCA[57]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 4 (left panel), illustrates only one class of observations with strong homogeneity and is well-modeled by PCA. This is commonly referred to as the asymmetric case. In Figure 4 (right panel), there are two homogeneous classes of observations, each separately modeled 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.

The SIMCA method is recommended to be use for the one class case, when we have one well-defined class of objects and all other objects are inhomogeneous, that is, asymmetric case. Another situation when SIMCA is the method of choice is for classification if no interpretation is needed.

Dumas et al.[58] have evaluated the anabolic treatment signature in cattle urine using NMR by an array of different statistical methods such as, ANOVA, LDA, and SIMCA classification for a two-class case. This is a typical case when SIMCA can be used, since only separation between cattle treated with steroids and nontreated cattle is initially required and less focus is on interpretation.

A few different examples wherein SIMCA have been applied are given in refs 59−61, However, if we have information from the study design regarding classes (sick/healthy, treated vs nontreated, etc.), it is recommended to use other supervised methods such as PLS-DA in the two-class cases (or multiple class cases) and/or preferably OPLS-DA[19] to facilitate the interpretation (Figure 5).

Partial Least-Squares (PLS) Method by Projections to Latent Structures. PLS[3,4,62,63] is a method commonly used where

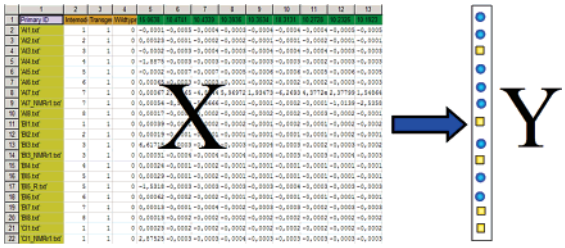

Figure 5. Class information can also be used to construct an additional matrix, hereinafter called the **Y** matrix, consisting of a discrete 'dummy' variable where [1]/[0] indicate the class belonging.

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. 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

Model of **X**: $\mathbf{X} = \mathrm{TP}^T + E$

Model of **Y**: $\mathbf{Y} = \mathrm{TC}^T + F$

PLS models are negatively affected by systematic variation in the **X** matrix that is not related to the **Y** matrix, 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.

Wang et al.[64,65] used LC/MS for profiling of the plasma phospholipids in Type 2 diabetes mellitus (DM2). By PLS-DA, it was possible not only to differentiate the DM2 from the controls group, but also to identify possible biomarkers. A number of examples applying PLS-DA have been published; for example, see refs 66−70.

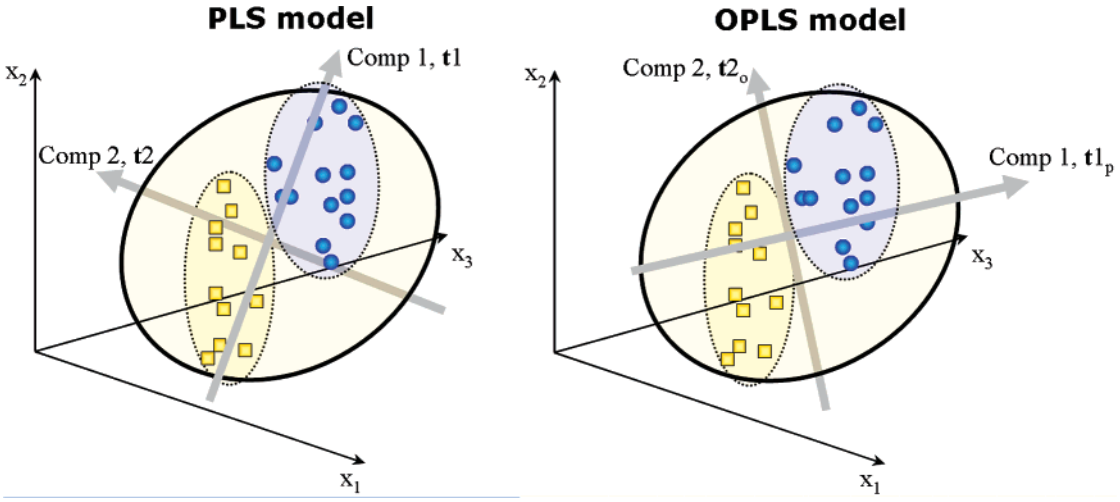

Figure 6. A geometrical illustration of the difference between the PLS-DA and OPLS-DA models. In the left panel, the PLS components cannot separate the between-class variation from the within-class variation, and the resulting PLS component loadings mixes both types of variations. In the right panel, the OPLS components are able to separate these two different variations. Component 1 ($\mathbf{t1}_p$) 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 ($\mathbf{t2}_o$) is the **Y**-orthogonal component and models the within group (within-class) variation.

In Brindle et al.,[71] PLS-DA was applied together with a multivariate preprocessing filter called orthogonal signal correction (OSC) for developing a diagnostic tool for predicting the severity of coronary heart disease based on NMR spectral profiles of human serum. The OSC filter removes the uncorrelated signals resulting in information of the within-class variation. Wagner et al. report the use of this, in a paper[72] wherein OSC was applied to investigate the background information, which was not due to the exposure of the compound acetaminophen. As stated by Wagner et al., the OSC component surprisingly provided an additional classification of male and females. These observations lead us to discuss the OPLS method.

The Orthogonal-PLS Method (OPLS). The OPLS[15] method is a recent modification of the PLS method.[3] 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.[15,19] The OPLS model comprises two modeled variations, the Y-predictive ($T_pP_p^T$) and the Y-orthogonal ($T_oP_o^T$) components. Only the Y-predictive variation is used for the modeling of **Y** ($T_pC_p^T$).

$$\text{Model of } \mathbf{X}: \quad \mathbf{X} = T_pP_p^T + T_oP_o^T + \mathbf{E}$$

$$\text{Model of } \mathbf{Y}: \quad \mathbf{Y} = T_pC_p^T + \mathbf{F}$$

E and **F** are the residual matrices of **X** and **Y**, respectively. OPLS can, analogously to PLS-DA, be used for discrimination (OPLS-DA); see, for instance, ref 19. In Figure 6, the advantages with OPLS-DA compared to PLS-DA are exemplified. The between-class variation and the within-class variation are separated by OPLS-DA but not by PLS-DA, and this facilitates the OPLS-DA model interpretation.

In Cloarec et al.,[73] a combined covariation and correlation-loading plot provided additional information on the physicochemical variations in the data. This was done by means of coloring each covariance loading by its correlation value to class separation.

Stella et al.[74] have illustrated the use OPLS for characterization of metabolic profile due to different diets and thereby identified difference in metabolic pattern between low-meat diet and vegetarian diet. This is the first systematic study reported on the dietary effects on the metabolism.

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.

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 are based on the expected or known pharmaco-kinetics of the expected effect. In other words, statistical experimental design 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 modeling 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.

Coen et al.[75] retrieved gene expression data and performed different types of NMR experiments of liver tissue from rats dosed with acetaminophen. The design study included 70 rats, 5 time points, and 4 different dose levels. Statistical analysis

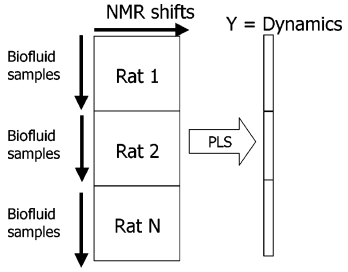

Figure 7. In batch modeling, the data is organised as an **X**-matrix containing blocks of rows where each block represents an object (e.g., a rat). Each row in a block represents the multivariate profile of an observation (e.g., the NMR spectral shifts) at a specific time point. The corresponding row 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.

and biochemical interpretation were performed using ANOVA and PCA modeling. In Smilde et al.,[76] the ANOVA-simultaneous, component analysis (ASCA) method is described as a direct generalization of analysis of variance (ANOVA) for univariate data to the multivariate case. ASCA was demonstrated on an intervention study examining the effect of vitamin C on the development of OA in guinea pigs.

Dieterle et al.[77] applied metabonomics to a compound ranking toxicity study where rat urine was collected and characterized using NMR spectroscopy. The design study included controls, multiple dose groups, and two time points. Arrays of classical end points were also assessed besides the NMR profile. Statistical modeling included PCA and PLS modeling.

Bollard et al.[78] presented an additional approach to dynamical studies by using the metabolic principal component trajectories to highlight the maximum effect. The data was scaled by scaling to maximum aligned and reduced trajectories (SMART), to remove differences in individual starting positions of the objects and varying magnitudes of effects and thereby facilitate the comparison between the objects. This type SMART-PCA is further illustrated in Keun et al.[79] Other examples of dynamic studies wherein PCA have been used or the evaluation of data have been handled as a two-class problem with PLS-DA are given in refs 80−84.

A different approach to handling dynamic metabolic data was presented by Antti et al.[85] who studied the time-dependent effect of hydrazine on the metabolic profiles for rats. Urine samples were collected from dosed rats and matched control animals at several time points up to 168 h post-dose. The samples were analyzed by NMR and evaluated by batch modeling.

Batch Modeling. Batch modeling[86] 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 modeling methods are used in metabonomic studies to model the time dependency or dynamics of biological processes,for example, the evolution 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 (Figure 7). Batch modeling is based on modeling 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 8. 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 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 modeling is based on the assumption that a control group of objects is followed over the same time period as the treated group.

A drawback with batch modeling is that all study objects must have a similar metabolic and response rate; we cannot have slow and fast responders in the same model.

Multi-"omics" Studies

Lindon et al.[87] discussed the huge problem of combining and using the data obtained in metabonomics studies from an array of analytical chemical methods as well as metabolic profiles from different compartments, and further on, to use these combined information for diagnosis, understanding physiological variation, drug therapy monitoring, as well as effect evaluation. If this can be handled, then metabonomics will be an integral part in the development of personalized medicine.

Attempts to approach this type of problems have been made by Klenø et al.,[88] by studying combined genomics, proteomics, and metabonomics data. Protein and endogenous metabolites were identified as altered in rats treated with hydrazine compared with untreated controls. The design study included a control group and two dosed groups with 10 rats in each. Statistical evaluation and interpretation were performed separately and provided an insight into the underlying biochemistry. Rantalainen et al.[89] demonstrated a novel approach using the OPLS method to integrate 2D-DIGE proteomic and NMR metabolic data from a human tumor xenograft mouse model of prostate cancer.

Yap et al.[90] used partial least-squares-based cross-correlation of the variation in the liver and plasma profiles and found that an increase in liver lipids correlated to a decrease in plasma lipids; this method will work nicely for two compartments. However, the complexity will rapidly increase with the number of compartments and "omic" data. This problem has been stressed by Craig et al.,[91] wherein the difficulties of combing and interpreting multiomic are discussed.

One way of approaching the multicompartment problem is presented by Martin et al.[92] They have applied H-PCA followed by OPLS-DA. This resulted in a model where both inter- and intracompartment covariance and correlations easily can be evaluated and thereby increase the understanding of biochemical mechanisms and relations between different compartments.

Hierarchical PCA. The idea behind hierarchical PCA is to block the variables to improve transparency and interpretability.[93−95] This method operates on two or more levels. 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 for two levels can be described as follows (see Figure 9): In the first step in this case is to divide the large matrix into conceptually meaningful blocks and make a separate Principal Component Analysis 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

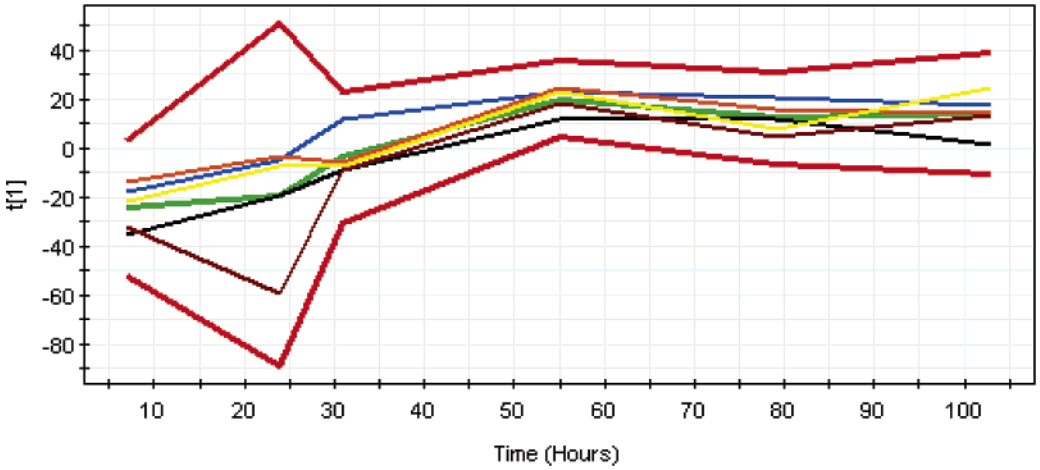

Figure 8. Batch control charts can be constructed from a PLS or OPLS batch model score vectors. The average score trajectory (for each component) with upper and lower control limits (based on standard deviations) indicates the normal dynamic trajectory for a batch. The control chart can be used for detecting deviations from normality.

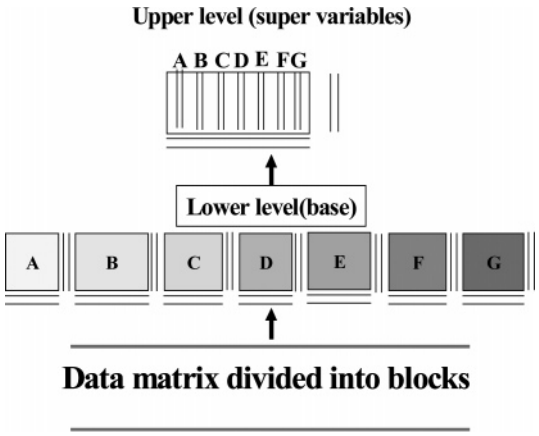

Figure 9. H-PCA is shown from the bottom to the top. At bottom of figure, the data matrix is divided into blocks. A separate PCA model is calculated for each block, and the PCA score components from each model are then combined to form a new matrix, summarizing all blocks. This new block of data is then analyzed by a PCA.

step, a PCA model fitted to this data and the hierarchical PCA model is established, see Figure 8.

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 are most 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.

Discussion And Future Remarks

In this review, we have provided an overview of how the underlying philosophy of chemometrics can be integrated throughout metabonomic studies. We have been able to illustrate each separate step with different examples from the literature showing the state of the art. The most common chemometrical tool used in the evaluation of a metabonomic study is PCA. PCA is always recommended as a starting point for analyzing multivariate data and will rapidly provide an overview of the information hidden in the data. Unfortunately, in a majority of the reviewed papers, the PCA method is the only tool applied. Often additional information can be extracted by using more advanced multivariate methods. In a few papers, PLS-DA and/or OPLS-DA have been used for modeling two classes of data to increase the class separation, simplify interpretation, and find potential biomarkers. For the two-class problem, OPLS-DA is recommended to obtain a clearer and more straightforward interpretation. It can also provide an understanding of the interclass variation.

A few papers also evaluate dynamic data, and one of the approaches used is batch modeling, a PLS-based method. A major drawback with this method is the assumption that all study objects have similar starting profiles and dynamics, for example, responding at the same metabolic rate to a treatment. This problem may be controlled by using a multivariate design for selecting the objects and thereby introduce a "controlled" biological variation.

There is a general lack in applying statistical experimental design (SED) to ensure balanced data and to have a defined experimental domain. The problems with multiomics data and combining data from different compartments have been discussed in a few papers. In one of the papers, the interesting multivariate approach by combining H-PCA with OPLS is suggested. This combination results in a straightforward way to handle the data as well as simplifying the interpretation, wherein different compartments are combined.

A future outlook for chemometrics in metabonomics is that the benefits of statistical experimental design in conjunction with more focused modeling methods such as PLS and OPLS become more widely known and applied to a much greater extent, not only for the two-class problems, but also for dynamic studies. However, it is likely to take some time until a fully integrated multivariate approach is published, based on the chemometric philosophy.

References

(1) Robertson, D. G.; Reily, M. D.; Baker, J. D. Metabonomics in preclinical drug development. *Expert Opin. Drug Metab. Toxicol.* **2005**. *1* (3), 363−376.
(2) Jackson J. E. *A Users Guide to Principal Components*; Wiley: New York, 1991.

(3) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III The Collinearity problem in linear regression. The partial least squares approach to generalized inverses. *SIAM J. Sci. Stat. Comput.* **1984**, *5* (3), 735−743.

(4) Wold, S.; Martens, H.; Wold, H. *Lecture Notes in Mathematics Proc. Conf. Matrix pencils, Piteå, Sweden*; Springer-Verlag: Heidelberg, 1983.

(5) Lundstedt, T.; Seifert, E.; Abramo, L.; Thelin, B.; Nyström, A.; Pettersen, J.; Bergman R. Experimental design and optimization. *Chemom. Intell. Lab. Syst.* **1998**, *42*, 3−40.

(6) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*; John Wiley & Sons: New York, 1978.

(7) Eriksson, L.; Johansson, E.; Kettaneh Wold, N.; Wikström, C.; Wold, S. *Design of Experiments—principles and Applications*, Umetrics AB, Umeå, Sweden, 1996.

(8) Antti, H.; Ebbels, T. M. D.; Keun, H. C.; Bollard, M. E.; Beckonert, O.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. Statistical experimental design and partial least squares regression analysis of biofluid metabonomic NMR and clinical chemistry data for screening of adverse drug effects. *Chemom. Intell. Lab. Syst.* **2004**, *73*, 139−149.

(9) Hotelling, H. The most predictable criterion. *J. Educ. Psychol.* **1935**, *26*, 139−142.

(10) Greenacre, M. J. *Theory and Applications of Correspondence Analysis*; Academic Press: London, 1984.

(11) Bishop, C. M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, U.K., 1996.

(12) Wythoff, B. J. Backpropagation neural networks—a tutorial. *Chemom. Intell. Lab. Syst.* **1993**, *18* (2), 115−155.

(13) Sivia, D. S. *Data Analysis: A Bayesian Tutorial*; Oxford University Press: Oxford, U.K., 1996.

(14) Rabiner, L. R.; Juang, B. H. An Introduction to Hidden Markov Models. *IEEE ASSP Mag.*, January, 1986.

(15) Trygg, J.; Wold S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **2002**, *16*, 119−128.

(16) Trygg, J. O2-PLS for qualitative and quantitative analysis in multivariate calibration. *J. Chemom.* **2002**, *16*, 283−293.

(17) Trygg, J.; Wold, S. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *J. Chemom.* **2003**, *17*, 53−64.

(18) 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. *Anal. Chem.* **2005**, *77* (2), 517−526.

(19) Bylesjö, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. OPLS Discriminant Analysis—Combining the strengths of PLS-DA and SIMCA classification. *J. Chemom.*, **2006**, in press.

(20) Kvalheim, O. M. The latent variable. *Chemom. Intell. Lab. Syst.* **1992**, *14*, 1−3.

(21) Wold, S.; Sjöström, M.; Carlson. R.; Lundstedt, T.; Hellberg, S.; Skageberg, B.; Wikström, C. Multivariate design. *Anal. Chim. Acta* **1986**, *17*, 191.

(22) Carlson, R.; Lundstedt, T. Scope of organic synthetic reactions. Multivariate methods for exploring the reaction space. An example of the Willgerodt-Kindler reaction. *Acta Chem. Scand. B* **1987**, *41*, 164.

(23) Carlson, R.; Lundstedt, T.; Albano, C. Screening of suitable solvents for organic synthesis, strategies for solvent selection. *Acta Chem. Scand, B* **1984**, *39*, 79.

(24) Sandberg, M.; Sjöström, M.; Jonsson, J. A multivariate characterization of tRNA nucleosides. *J. Chemom.* **1996**, *10*, 493−508.

(25) Oprea, T. I.; Gottfries, J. Chemography: The art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3* (2), 157−166.

(26) deAguiar, P. F.; Bourguignon, B.; Khots, M.; Massart, D. L.; PhanThanLuu R. D-optimal designs. *Chemom. Intell. Lab. Syst.* **1995**, *30* (2), 199−210.

(27) The Standard Metabolic Reporting Structure, Version 2.3, http://www.smrsgroup.org/, Jan uary 13, 2006.

(28) 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. *Anal. Biochem.* **2004**, *331*, 283−295.

(29) 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. *Anal. Chem.* **2005**, *77*, 8086−8094.

(30) Dumas, M. E.; Maibaum, E. C.; Teague, C.; Ueshima, H.; Zhou, B.; Lindon, J. C.; Nicholson, J. K.; Stamler, J.; Elliott, P.; Chan, Q.; Holmes, E. Assessment of analytical reproducibility of 1H NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP Study. *Anal. Chem.* **2006**, *78*, 2199−208.

(31) Dumas, M. E.; Canlet, C.; Debrauwer, L.; Martin, P.; Paris, A. Selection of biomarkers by a multivariate statistical processing of composite metabonomic data sets using multiple factor analysis. *J. Proteome Res.* **2005**, *4*, 1485−1492.

(32) 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.* **2004**, *76*, 1738−1745.

(33) 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* **2005**, *130*, 701−707.

(34) 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.* **1999**, *13*, 279−284.

(35) Shen, H. L.; Grung, B.; Kvalheim, O. M.; Eide. I. Automated curve resolution applied to data from multi-detection instruments. *Anal. Chim. Acta* **2001**, *446* (1−2), 313−328.

(36) Jonsson, P.; Sjövik Johansson, E.; Wuolikainen, A.; Lindberg, J.; Schuppe-Koistinen, I.; Kusano, M.; Sjöström, M.; Trygg, J.; Moritz, T.; Antti, H. Predictive metabolite profiling applying hierarchical multivariate curve resolution to GC-MS data-A potential tool for multi-parametric diagnosis. *J. Proteome Res.* **2006**, *5* (6), 1407−1414.

(37) Torgrip, R. J. O.; Aberg. M.; Karlberg, B.; Jacobsson, S. P. Peak alignment using reduced set mapping. *J. Chemom.***2003**, *17* (11), 573−582.

(38) 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. *Chemom. Intell. Lab. Syst.* **1993**, *21*, 2−3, 249−258.

(39) Holmes, E.; Nicholson, J. K.; Nicholls, A. W.; Lindon, J. C.; Connor, S. C.; Polley, S.; Connelly, J. The identification of novel biomarkers of renal toxicity using automatic data reduction techniques and PCA of proton NMR spectra of urine. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 245−255.

(40) Stoyanova, R.; Nicholls, A. W.; Nicholson, J. K.; Lindon, J. C.; Brown, T. R. Automatic alignment of individual peaks in large high-resolution spectral data sets. *J. Magn. Reson.* **2004**, *170*, 329−35.

(41) Forshed, J. R.; Torgrip, J. O.; Aberg, K. M.; Karlberg, B.; Lindberg, J.; Jacobsson, S. P. A comparison of methods for alignment of NMR peaks in the context of cluster analysis. *J. Pharm. Biomed. Anal.* **2005**, *38*, 824−832.

(42) Forshed, J.; Schuppe-Koistinen, I.; Jacobsson, S. P. Peak alignment of NMR signals by means of a genetic algorithm. *Anal. Chim. Acta* **2003**, *487*, 189−199.

(43) Lee, G. C.; Woodruff, D. L. Beam search for peak alignment of NMR signals. *Anal. Chim. Acta* **2004**, *513*, 413−416.

(44) Hotelling, H. The generalization of Student's ratio. *Ann. Math. Stat.* **1931**, *2*, 360−378.

(45) Eriksson, L.; Johansson, E.; Kettaneh Wold, N.; Wold, S. *Multi and Megavariate Data Analysis*; Umetrics AB, Umeå, Sweden, 2001.

(46) Miller, P.; Swanson, R. E.; Heckler, C. E. Contribution plots: A missing link in multivariate quality control. *Appl. Math. Comp. Sci.* **1998**, *8* (4), 775−792.

(47) Akira, K.; Imachi, M.; Hashimoto, T. Investigations into biochemical changes of genetic hypertensive rats using 1H nuclear magnetic resonance-based metabonomics. *Hypertens. Res.* **2005**, *28*, 425−430.

(48) Yang, J.; Xu, G.; Zheng, Y.; Kong, H.; Pang, T.; Lv, S.; Yang, Q. Diagnosis of liver cancer using HPLC-based metabonomics avoiding false-positive result from hepatitis and hepatocirrhosis diseases. *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.* **2004**, *813*, 59−65.

(49) Lenz, E. M.; Bright, J.; Wilson, I. D.; Morgan, S. R.; Nash, A. F. A 1H NMR-based metabonomic study of urine and plasma samples obtained from healthy human subjects. *J. Pharm. Biomed. Anal.* **2003**, *33*, 1103−1115.

(50) Wang, Y.; Bollard, M. E.; Keun, H.; Antti, H.; Beckonert, O.; Ebbels, T. M.; Lindon, J. C.; Holmes, E.; Tang, H.; Nicholson, J. K. Spectral editing and pattern recognition methods applied to high-resolution magic-angle spinning 1H nuclear magnetic resonance spectroscopy of liver tissues. *Anal. Biochem.* **2003**, *323*, 26−32.
(51) Plumb, R. S.; Stumpf, C. L.; Gorenstein, M. V.; Castro-Perez, J. M.; Dear, G. J.; Anthony, M.; Sweatman, B. C.; Connor, S. C.; Haselden, J. N. Metabonomics: The use of electrospray mass spectrometry coupled to reversed-phase liquid chromatography shows potential for the screening of rat urine in drug development. *Rapid. Commun. Mass. Spectrom.* **2002**, *16*, 1991−1996.
(52) 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 recognition technology for rapid in vivo screening of liver and kidney toxicants. *Toxicol. Sci.* **2000**, *57*, 326−337.
(53) Kim, S. W.; Ban, S. H.; Ahn, C. Y.; Oh, H. M.; Chung, H.; Cho, S. H.; Park, Y. M.; Liu, J. R. Taxonomic discrimination of cyanobacteria by metabolic fingerprinting using proton nuclear magnetic resonance spectra and multivariate statistical analysis. *J. Plant Biol.* **2006**, *49*, 271−275.
(54) Rasmussen, B.; Cloarec, O.; Tang, H. R.; Staerk, D.; Jaroszewski, J. W. Multivariate analysis of integrated and full-resolution H-1-NMR spectral data from complex pharmaceutical preparations: St. John's wort. *Planta Med.* **2006**, *72*, 556−563.
(55) Chen, H. W.; Pan, Z. Z.; Talaty, N.; Raftery, D.; Cooks, R. G. Combining desorption electrospray ionization mass spectrometry and nuclear magnetic resonance for differential metabolomics without sample preparation. *Rapid. Commun. Mass Spectrom.* **2006**, *20*, 1577−1584.
(56) Halouska, S.; Powers, R. Negative impact of noise on the principal component analysis of NMR data. *J. Magn. Reson.* **2006**, *178*, 88−95.
(57) Wold, S. Pattern recognition by means of disjoint principal components models. *Pattern Recognit.* **1976**, *8*, 127−139.
(58) Dumas, M. E.; Canlet, C.; Vercauteren, J.; Andre, F.; Paris, A. Homeostatic signature of anabolic steroids in cattle using H-1-C-13 HMBC NMR metabonomics. *J. Proteome Res.* **2005**, *4*, 1493−1502.
(59) Odunsi, K. R.; Wollman, M.; Ambrosone, C. B.; Hutson, A.; McCann, S. E.; Tammela, J.; Geisler, J. P.; Miller, G.; Sellers, T.; Cliby, W.; Qian, F.; Keitz, B.; Intengan, M.; Lele, S.; Alderfer, J. L. Detection of epithelial ovarian cancer using 1H-NMR-based metabonomics. *Int. J. Cancer* **2005**, *113*, 782−788.
(60) Holmes, E.; Nicholls, A. W.; Lindon, J. C.; Connor, S. C.; Connelly, J. C.; Haselden, J. N.; Damment, S. J.; Spraul, M.; Neidig, P.; Nicholson, J. K. Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chem. Res. Toxicol.* **2000**, *13*, 471−478.
(61) McKee, C. L. G.; Wilson, I. D.; Nicholson, J. K. Metabolic phenotyping of nude and normal (Alpk : ApfCD, C57BL10J) mice. *J. Proteome Res.* **2006**, *5*, 378−384.
(62) Wold, S.; Eriksson, L.; Sjöström, M. PLS in Chemistry. *Encyclopedia of Computational Chemistry*; Schleyer, P. V. R., Ed.; John Wiley & Sons: New York, 1998; pp2006−2016.
(63) Wold, S.; Albano, C.; Dunn, W. J., III; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johansson. E.; Lindberg, W.; Sjöström, M. In *Multivariate Data Analysis in Chemistry*; NATO ASI Series C 138, D; Reidel Publ. Co.: Dordrecht, Holland, 1984.
(64) Wang, C. H.; Kong, W.; Guan, Y. F.; Yang, J.; Gu, J. R.; Yang, S. L.; Xu, G. W. Plasma phospholipid metabolic profiling and biomarkers of type 2 diabetes mellitus based on high-performance liquid chromatography/electrospray mass spectrometry and multivariate statistical analysis. *Anal. Chem.* **2005**, *77*, 4108−4116.
(65) Wang, C.; Kong, H.; Guan, Y.; Yang, J.; Gu, J.; Yang, S.; Xu, G. Plasma phospholipid metabolic profiling and biomarkers of type 2 diabetes mellitus based on high-performance liquid chromatography/electrospray mass spectrometry and multivariate statistical analysis. *Anal. Chem.* **2005**, *77*, 4108−4116.
(66) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Metabolite projection analysis for fast identification of metabolites in metabonomics. Application in an amiodarone study. *Anal. Chem.* **2006**, *78*, 3551−3561.
(67) Yin, P.; Zhao, X.; Li, Q.; Wang, J.; Li, J.; Xu, G. Metabonomics study of intestinal fistulas based on ultraperformance liquid chromatography coupled with Q-TOF mass spectrometry (UPLC/Q-TOF MS). *J. Proteome Res.* **2006**, *5*, 2135−2143.
(68) Ramadan, Z.; Jacobs, D.; Grigorov, M.; Kochhar, S. Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms. *Talanta* **2006**, *68*, 1683−1691.
(69) Constantinou, M. A.; Papakonstantinou, E.; Spraul, M.; Sevastiadou, S.; Costalos, C.; Koupparis, M. A.; Shulpis, K.; Tsantili-Kakoulidou, A.; Mikros, E. H-1 NMR-based metabonomics for the diagnosis of inborn errors of metabolism in urine. *Anal. Chim. Acta* **2005**, *542*, 169−177.
(70) Yang, J.; Zhao, X.; Liu, X.; Wang, C.; Gao, P.; Wang, J.; Li, L.; Gu, J.; Yang, S.; Xu, G. High performance liquid chromatography-mass spectrometry for metabonomics: potential biomarkers for acute deterioration of liver function in chronic hepatitis B. *J. Proteome Res.* **2006**, *5*, 554−561.
(71) Brindle, J. T.; Antti, H.; Holmes, E.; Tranter, G.; Nicholson, J. K.; Bethell, H. W.; Clarke, S.; Schofield, P. M.; McKilligin, E.; Mosedale, D. E.; Grainger, D. J. Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nat. Med.* **2002**, *8*, 1439−1444.
(72) Wagner, S.; Scholz, K.; Donegan, M.; Burton, L.; Wingate, J.; Volkel, W. Metabonomics and biomarker discovery: LC-MS metabolic profiling and constant neutral loss scanning combined with multivariate data analysis for mercapturic acid analysis. *Anal. Chem.* **2006**, *78*, 1296−1305.
(73) Cloarec, O. M.; Dumas, E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic H-1 NMR data sets. *Anal. Chem.* **2005**, *77*, 1282−1289.
(74) Stella, C.; Beckwith-Hall, B.; Cloarec, O.; Holmes, E.; Lindon, J. C.; Powell, J.; van der Ouderaa, F.; Bingham, S.; Cross, A. J.; Nicholson, J. K. Susceptibility of human metabolic phenotypes to dietary modulation. *J. Proteome Res.* **2006**, *5*, 2780−2788.
(75) Coen, M.; Ruepp, S. U.; Lindon, J. C.; Nicholson, J. K.; Pognan, F.; Lenz, E. M.; Wilson, I. D. Integrated application of transcriptomics and metabonomics yields new insight into the toxicity due to paracetamol in the mouse. *J. Pharm. Biomed. Anal.* **2004**, *35*, 93−105.
(76) Smilde, A. K.; Jansen, J. J.; Hoefsloot, H. C. J.; Lamers, R. J. A. N.; van der Greef, J.; Timmerman, M. E. ANOVA-simultaneous, component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics* **2005**, *21*, 3043−3048.
(77) Dieterle, F.; Schlotterbeck, G. T.; Ross, A.; Niederhauser, U.; Senn, H. Application of metabonomics in a compound ranking study in early drug development revealing drug-induced excretion of choline into urine. *Chem. Res. Toxicol.* **2006**, *19*, 1175−1181.
(78) Bollard, M. E.; Keun, H. C.; Beckonert, O.; Ebbels, T. M.; Antti, H.; Nicholls, A. W.; Shockcor, J. P.; Cantor, G. H.; Stevens, G.; Lindon, J. C.; Holmes, E.; Nicholson, J. K. Comparative metabonomics of differential hydrazine toxicity in the rat and mouse. *Toxicol. Appl. Pharmacol.* **2005**, *204*, 135−151.
(79) Keun, H. C.; Ebbels, T. M.; Bollard, M. E.; Beckonert, O.; Antti, H.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. *Chem. Res. Toxicol.* **2004**, *17*, 579−587.
(80) Ishihara, K.; Katsutani, N.; Aoki, T. A metabonomics study of the hepatotoxicants galactosamine, methylene dianiline and clofibrate in rats. *Basic Clin. Pharmacol. Toxicol.* **2006**, *99*, 251−260.
(81) Williams, R. E.; Lenz, E. M.; Lowden, J. S.; Rantalainen, M.; Wilson, I. D. The metabonomics of aging and development in the rat: an investigation into the effect of age on the profile of endogenous metabolites in the urine of male rats using 1H NMR and HPLC-TOF MS. *Mol. Biosyst.* **2005**, *1*, 166−175.
(82) Williams, R. E.; Lenz, E. M.; Rantalainen, M.; Wilson, I. D. The comparative metabonomics of age-related changes in the urinary composition of male Wistar-derived and Zucker (fa/fa) obese rats. *Mol. Biosyst.* **2006**, *2*, 193−202.
(83) Schnackenberg, L. K.; Jones, R. C.; Thyparambil, S.; Taylor, J. T.; Han, T.; Tong, W.; Hansen, D. K.; Fuscoe, J. C.; Edmondson, R. D.; Beger, R. D.; Dragan, Y. P. An integrated study of acute effects of valproic acid in the liver using metabonomics, proteomics, and transcriptomics platforms. *OMICS* **2006**, *10*, 1−14.
(84) Bollard, M. E.; Keun, H. C.; Beckonert, O.; Ebbels, T. M. D.; Antti, H.; Nicholls, A. W.; Shockcor, J. P.; Cantor, G. H.; Stevens, G.; Lindon, J. C.; Holmes, E.; Nicholson, J. K. Comparative metabonomics of differential hydrazine toxicity in the rat and mouse. *Toxicol. Appl. Pharmacol.* **2005**, *204*, 135−151.
(85) Antti, H.; Bollard, M. E.; Ebbels, T.; Keun, H.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. Batch statistical processing of H-1 NMR-derived urinary spectral data. *J. Chemom.* **2002**, *16*, 461−468.
(86) Wold, S.; Kettaneh, N.; Friden, H.; Holmberg, A. Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 331−340.

(87) Lindon, J. C.; Holmes, E.; Nicholson, J. K. Metabonomics: Systems biology in pharmaceutical research and development. *Curr. Opin. Mol. Ther.* **2004**, *6*, 265−272.
(88) Kleno, T. G.; Kiehr, B.; Baunsgaard, D.; Sidelmann, U. G. Combination of 'omics' data to investigate the mechanism(s) of hydrazine-induced hepatotoxicity in rats and to identify potential biomarkers. *Biomarkers* **2004**, *9*, 116−138.
(89) Rantalainen, M.; Cloarec, O.; Beckonert, O.; Wilson, I. D.; Jackson, D.; Tonge, R.; Rowlinson, R.; Rayner, S.; Nickson, J.; Wilkinson, R. W.; Mills, J. D,; Trygg, J.; Nicholson, J. K.; Holmes, E. Statistically integrated metabonomic-proteomic studies on a human prostate cancer xenograft model in mice. *J. Proteome Res.* **2006**, *5* (10), 2642−2655.
(90) Yap, I. K.; Clayton, T. A.; Tang, H.; Everett, J. R.; Hanton, G.; J. Provost, P.; Le Net, J. L.; Charuel, C.; Lindon, J. C.; Nicholson, J. K. An integrated metabonomic approach to describe temporal metabolic disregulation induced in the rat by the model hepatotoxin allyl formate. *J. Proteome. Res.* **2006**, *5*, 2675−2684.
(91) Craig, A.; Sidaway, J.; Holmes, E.; Orton, T.; Jackson, D.; Rowlinson, R.; Nickson, J.; Tonge, R.; Wilson, I.; Nicholson, J. Systems toxicology: integrated genomic, proteomic and metabonomic analysis of methapyrilene induced hepatotoxicity in the rat. *J. Proteome Res.* **2006**, *5*, 1586−1601.
(92) Martin, F. P. J.; Wang, Y.; Yap, I. K. S.; Lundstedt, T.; Lek, P.; Lindon, J. C.; Sprenger, N.; Kochhar, S.; Fay, L. B.; Holmes, E.; Nicholson, J. K. NMR and UPLC-MS based multi-compartment metabonomic investigation of the contribution of different dietary probiotics to host metabolism. *J. Proteome Res.*, to be submitted for publication.
(93) Wold, S.; Kettaneh, N.; Tjessem, K. Hierarchical multiblock, PLS and PC models for easier model interpretation and as an alternative to variable selection. *J. Chemom.* **1996**, *10* (5−6), 463−482.
(94) Eriksson, L.; Johansson, E.; Lindgren. F.; Sjöström, M.; Wold, S. Megavariate analysis of hierarchical QSAR data. *J. Comput.−Aided Mol. Des.* **2002**, *16*, 711−726.
(95) Gunnarsson. I.; Andersson, P. M.; Wikberg, J.; Lundstedt, T. Multivariate analysis of G protein-coupled receptors. *J. Chemom.* **2003**, *17*, 82−92.

PR060594Q