ASCA

Jansen, J.J.

Publication date
2005

Document Version
Final published version

Link to publication

Citation for published version (APA):
Jansen, J. J. (2005). *ASCA*.

ASCA

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus

prof. mr. P.F. van der Heijden

ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit

op dinsdag 22 november 2005, te 14:00 uur

door **Jeroen Jasper Jansen**

geboren te Zaanstad

Promotiecommissie:

Promotor:
Prof. Dr. A.K. Smilde

Co-promotores:
Prof. Dr. J. van der Greef
Dr. H.C.J. Hoefsloot

Overige leden:
Dr. M.E. Timmerman
Prof. Dr. S. Brul
Prof. Dr. C.G. de Koster
Prof. Dr. O. Fiehn
Prof. Dr. Thomas Hankemeier

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Table of Contents

1 Introduction

This thesis is about the analysis of time-resolved metabolomics data. For this analysis, models are used that employ *a priori* knowledge about the experiment and the data. First ‘metabolomics’ is explained. Subsequently the use of *a priori* knowledge in multivariate data analysis is explained. Finally the multivariate data analysis models that are described in this thesis are briefly discussed.

1.1 Metabolomics

Any organism can be seen as consisting of a series of highly interlinked and complex system parts. In Systems Biology the properties of these parts and the relations between them are studied from a holistic viewpoint. Several ‘omics’-techniques have been developed for this: for example DNA transcription is monitored by transcriptomics and the proteins that are present in an organism are analyzed using proteomics (1).

Metabolomics is the ‘omics’ method in which the metabolism of an organism is analyzed, based on the measurement of the concentration of all (or most) of its metabolites. Thereby information is collected about the whole metabolic system of the organism. This distinguishes metabolomics from more conventional approaches for investigating the metabolism that focus on a specific metabolite or a limited set of compounds (2).

In this thesis only case studies of metabolomics on mammals are presented, in which the composition of their body fluids is measured to determine the ‘status’ of their metabolism. A lot of work has been done in analytical method development for mammalian metabolomics. The use of various body fluids has been proposed (3-5) and different analytical platforms have been suggested for the analysis of their metabolite composition (6). Urine is used in all applications described in this thesis. Its metabolite composition has been measured using proton Nuclear Magnetic Resonance (^{1}H-NMR) spectroscopy. This approach, which is more focused on a rapid screening of the metabolism, is often referred to as metabolic fingerprinting (7) or metabonomics (8). Although the applications in this thesis are limited to this fingerprinting, the ideas presented here can be applied in all fields of metabolomics.

Several overview articles about metabolomics are available. A general overview about the history of metabolomics and its current challenges is given by Van Der Greef *et al.* (9). A paper specifically focused on metabonomics is written by Nicholson *et al.* (10). Also the articles by Fiehn (2) and by Dunn *et al.* (6) that were mentioned earlier give a good overview about the current possibilities and limitations of metabolomics.

1.2 Black, white and grey models

The NMR spectra collected from a metabolic fingerprinting experiment describe the concentrations of the metabolites in a urine sample. These concentrations contain the information about the status of metabolism. The collected spectra are information-rich: a visual inspection will not reveal all information about the metabolism contained within this data. To extract this information, multivariate data analysis techniques are required (11, 12). Usually these data analysis methods are similar or equal to methods from chemometrics: the field that deals with analysis of data collected from experiments in chemistry (13, 14).

The selection of a data analysis method for a specific experimental question and corresponding dataset is dependent on several considerations. One of these considerations is the amount of *a priori* information that should be put into the model. Generally more information than the collected data itself is available from an experiment, such as information about the experimental design or information about the mechanisms underlying the studied phenomenon. The more of this information is used to construct the model, the stricter it becomes. However, the information introduced to the model usually simplifies its interpretation.

Models that use no *a priori* information at all are referred to as ‘soft’ or ‘black’ models. Models that are derived from physical or chemical laws (for example differential equations that describe pathways present in an organism) are referred to as ‘hard’ or ‘white’ models. The drawback of black models is that they are generally difficult to interpret in terms of the natural factors underlying the variation of the data and the drawback of white models is that quite often the physical or chemical background of the observed system is not completely known.

The models that are discussed in this thesis are ‘grey’ models (15). These models are black, but the available *a priori* (‘white’) information is used to improve their interpretability. Grey models are not as strict as white models, but they are generally better interpretable than black models. A visual depiction of these different types of models is given in Figure 1.

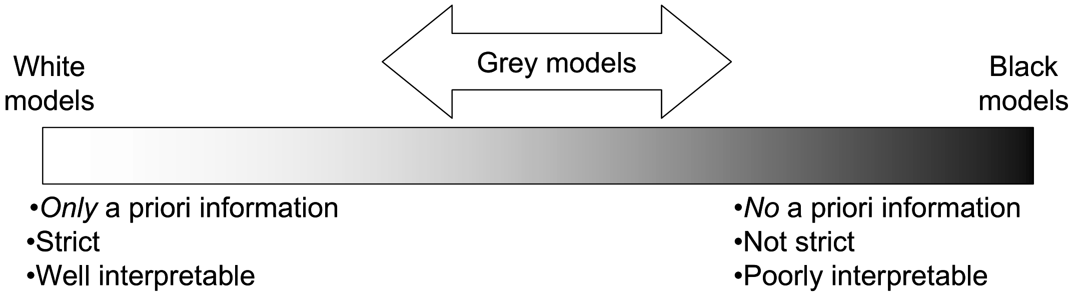

Figure 1 Distinction between white, grey and black models

1.3 Multivariate Data analysis of metabolomics data

PCA is the most widely used method for the exploratory data analysis of time-resolved metabolomics data (e.g. (16-18)). It makes no assumptions about the variation in the data and therefore it is a black model. Due to the limited interpretability of PCA the mechanisms underlying changes in metabolism often remain elusive. The need for models with increased interpretability is certainly present.

In this thesis, two ‘grey’ models are used for the analysis of time-resolved metabolomics data. In Chapter 2 the *a priori* information that is used in the data analysis is knowledge about the measurement error. In the remainder of the thesis the design of the experiment is used as prior information for fitting a component model.

In Chapter 2, Weighted PCA (WPCA) is described (19-21). This method can be used when the error of a signal is dependent on the magnitude of this signal (heteroscedastic), which cannot be handled well by PCA. The chapter describes the process from obtaining the *a priori* information from the replicate measurements performed in the experiment, to the modeling of the measurement error and the final implementation of the information into the data analysis.

Chapters 3 to 6 cover ANOVA-SCA (ASCA). This method is developed because PCA cannot distinguish between different factors and interactions in the experimental design. Therefore it is combined with Analysis of Variance (ANOVA), which is generally used for the analysis of a designed experiment. Thereby ANOVA-SCA (ASCA) is obtained, which is a novel multivariate data analysis tool that takes the experimental design into account.

A precursor of ASCA has been proposed first by Timmerman (22). This method, developed for the multivariate analysis of hierarchically organized (*multilevel*) data. is called Multilevel Simultaneous Component Analysis (MSCA) (22). MSCA is used for the analysis of a time-resolved metabolomics dataset in Chapter 3 of this thesis. In Chapter 4 ASCA is described and applied to a disease intervention study of the effect of Vitamin C to the development of osteoarthritis in guinea pigs. Chapter 5 describes the applicability of the combination of ASCA and metabolomics in a case study from systems biology. The experimental question in this study is to determine the *homeostatic capacity* of rats for bromobenzene, a model toxic compound that acts on the liver. It is shown that this question can be indeed answered with the constructed ASCA model. In Chapter 6 the mathematical framework behind ASCA is explained. Also relations of ASCA to other methods are described in this chapter, as well as possible extensions of the method.

The chapters in this thesis are each based on a finished manuscript and therefore can be read independently of each other.

1.4 References

(1) Morel, N., Holland, J.M., Van der Greef, J., Marple, E.W., Clish, C.B., Loscalzo, J. and Naylor, S., Primer on Medical Genomics Part XIV: Introduction to Systems Biology—A New Approach to Understanding Disease and Treatment. *Mayo Clinics Proceedings*, 2004; **79**: 651-658

(2) Fiehn, O., Metabolomics-the link between genotypes and phenotypes. *Plant Molecular Biology*, 2002; **48**: 155-171

(3) Nicholson, J.K., Buckingham, M.J. and Sadler, P.J., High resolution 1 H-NMR studies of vertebrate blood and plasma. *Biochemistry Journal*, 1983; **211**: 605

(4) Tate, A.R., Stephen, J.P.D. and John, C.L., Investigation of the metabolite variation in control rat urine using 1 H NMR Spectroscopy. *Anal. Biochem.*, 2001; **291**: 17

(5) Satake, M., Dmochowska, B., Nishikawa, Y., Madaj, J., Xue, J., Guo, Z., Reddy, D.V., Rinaldi, P.L. and Monnier, V.M., Vitamin C Metabolomic Mapping in the Lens with 6-Deoxy-6-fluoro-ascorbic Acid and High-Resolution 19F-NMR Spectroscopy. *Invest. Ophthalmol. Vis. Sci.*, 2003; **44**: 2047-2058

(6) Dunn, W.B., Bailey, N.J. and Johnson, H.E., Measuring the metabolome: current analytical technologies. *The Analyst*, 2005; **130**: 606-625

(7) Lamers, R.-J.A.N., DeGroot, J., Spies-Faber, E.J., Jellema, R.H., Kraus, V.B., Verzijl, N., TeKoppele, J.M., Spijksma, G.K., Vogels, J.T.W.E., van der Greef, J. and van Nesselrooij, J.H.J., Identification of Disease- and Nutrient-Related Metabolic Fingerprints in Osteoarthritic Guinea Pigs. *J. Nutr.*, 2003; **133**: 1776-1780

(8) Nicholson, J.K., Lindon, J.C. and 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**: 1181

(9) van der Greef, J., Davidov, E., Verheij, E.R., van der Heijden, R., Adourian, A.S., Oresic, M., Marple, E.W., Naylor, S., Harrigan, G.G. and Goodacre, R., *The role of metabolomics in Systems Biology* in Metabolic Profiling: Its role in Biomarker Discovery and Gene Function Analysis Kluwer Academic Publishers, Boston/Dordrecht/London, 2003

(10) Nicholson, J.K., Connelly, J.C., Lindon, J.C. and Holmes, E., Metabonomics: a platform for studying drug toxicity and gene function. *Nature Reviews Drug Discovery*, 2002; **1**: 153

(11) Holmes, E., Nicholls, A.W., Lindon, J.C., Connor, S.C., Connelly, J.C., Haselden, J.N., Damment, S.J.P., Spraul, M., Neidig, P. and Nicholson, J.K., Chemometric models for toxicity classification based on NMR spectra of Biofluids. 2000; **13**: 471

(12) Lindon, J.C., Holmes, E. and Nicholson, J.K., Pattern Recognition methods and applications in biomedical magnetic resonance. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 2000; **39**: 1-40

(13) Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., De Jong, S., Lewi, P.J. and Smeyers-Verbeke, J., *Handbook of Chemometrics and Qualimetrics: Part A*. 20A, Elsevier, Amsterdam, 1997

(14) Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., De Jong, S., Lewi, P.J. and Smeyers-Verbeke, J., *Handbook of Chemometrics and Qualimetrics: Part B*. Elsevier Science, Amsterdam, 1998

(15) Van Sprang, E.N.M., Ramaker, H.J., Westerhuis, J.A., Smilde, A.K., Gurden, S.P. and Wienke, D., Near-infrared spectroscopic monitoring of a series of industrial batch processes using a bilinear grey model. *Applied Spectroscopy*, 2003; **57**: 1007-1019

(16) Beckwith-Hall, B.M., Nicholson, J.K., Nicholls, A.W., Foxall, P.J.D., Lindon, J.C., Connor, S.C., Abdi, M. and Holmes, E., Nuclear Magnetic Resonance Spectroscopic and Principal Components Analysis Investigations into biochemical effects of three model hepatotoxins. 1998; **11**: 260

(17) Potts, B.C.M., Deese, A.J., Stevens, G.J., Reilly, M.D., Robertson, D.G. and Theiss, J., NMR of biofluids and pattern recognition: assessing the impact of NMR parameters on the principal component analysis of urine from rat and mouse. *Journal of Pharmaceutical and Biomedical Analysis*, 2001; **26**: 463

(18) Holmes, E. and Antti, H., Chemometric contributions to the evolution of metabonomics: mathematical solutions to characterising and interpreting complex biological NMR Spectra,

(19) Kiers, H.A.L., Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, 1997; **62**: 251

(20) Bro, R., Smilde, A.K. and Sidiropoulos, N.D., Maximum Likelihood fitting using ordinary least squares algorithms. *Journal of Chemometrics*, 2002; **16**: 387

(21) Wentzell, P.D., Andrews, D.T., Hamilton, D.C., Faber, K. and Kowalski, B.R., Maximum likelihood principal component analysis. *Journal of Chemometrics*, 1997; **11**: 339-366

(22) Timmerman, M.E., Multilevel Component Analysis. *British Journal of Mathematical and Statistical Psychology*; In Press

2 Analysis of Metabolomics Data using Weighted PCA

2.1 Introduction

In genomics and systems biology a range of new methods has been developed that investigate cell states on different aggregation levels. These methods form the basis for the analysis of the processes that lead from the DNA-code to phenotypic changes in an organism. Three of these methods are transcriptomics, proteomics and metabolomics; all three generating large amounts of data. Contained within this data is the information about the organism that is investigated. To obtain this information, various types of data analysis methods are used.

Metabolomics investigates the metabolism of an organism. Specifically, the effect of certain influences (e.g. diet, toxic stress or disease) on metabolism is the focus of research. In metabolomics the chemical composition of cells, tissue or body fluids is studied and within Life Sciences a major emphasis is on studying effects on the metabolite pattern in the development of diseases (biomarker or disease research) or the effect of drugs on this pattern, so called drug response profiling. Although the term metabolomics is from recent years, the approach of metabolite fingerprinting and multivariate statistics has its origin in the seventies for the fingerprinting part and in the early eighties on the combined approaches (van der Greef, Davidov et al., 2003). In the development of these strategies focusing on disease biomarker patterns it became clear that biomarker patterns of living systems on a single time point or evaluated over different objects generate important information but do not capture the dynamics of a system and clear evidence has been generated that deregulation of the dynamics of a system can be the onset of disease development (dynamic disease concept, (Glass and Mackey, 1988)). Taking into account time information was explored by investigating pre-menstrual syndrome (PMS), by applying metabolomics while using the information on the menstrual cycle enabling the detection of trend-specific changes in PMS (Tas, van den Berg et al., 1989). The key to investigate dynamic phenomena is the analysis of time series and a better understanding of "normality".

Using non-invasive techniques like the analysis of urine (urinalysis) based on metabolomics is an attractive approach and has been used in many studies since the early onset in clinical chemistry related profiling.

A technique that is suitable for the analysis of the chemical composition of urine is ^{1}H-Nuclear Magnetic Resonance (NMR) spectroscopy (Holmes, Foxall *et al.*, 1994). Spectra of urine obtained by NMR-spectroscopy are complicated and have a high information density. The desired information can be extracted from these spectra using multivariate statistical methods commonly used in pattern recognition and chemometrics (Holmes and Antti, 2002).

The unsupervised analysis of the variation in the data is important in metabolomics. Principal Component Analysis (PCA) (Jolliffe, 2002) is a method that is often used for this. PCA is also applied in transcriptomics, proteomics and plant metabolomics (Heijne, Stierum *et al.*, 2003; Taylor, King *et al.*, 2002). Methods like PCA give a simplified representation of the information that is contained in NMR-spectra.

Weighted PCA (WPCA) (Kiers, 1997) is a method for unsupervised data analysis that is related to PCA. In WPCA each element of the data can be given a corresponding 'weight'. These weights can be defined using different sources of information. Examples of information that can be introduced into the data analysis by using WPCA are scaling constants concerning the relative importance of variables or samples (Bro and Smilde, 2003). Also autoscaling, sometimes performed as a data pre-processing method in PCA, can be seen as a specific case of WPCA (Paatero and Tapper, 1993). Missing values in the data can be accommodated by defining zero weights for these values and can also fall in the framework of WPCA (Andrews and Wentzell, 1997). Information about the error in the measurements (Wentzell, Andrews *et al.*, 1997) can be included in the data analysis to obtain a Maximum Likelihood PCA-model when errors are non-uniform. The definition of the weights can be generalised to include more forms of problem-specific *a priori* information about the data (Bro, Smilde *et al.*, 2002). The broad range of mentioned applications shows that WPCA is a generic bioinformatics tool to introduce *a priori* information into the data-analysis of e.g. metabolomics data.

The use of additional information in an analysis of metabolomics data is illustrated by the following application. The urine of healthy rhesus monkeys has been sampled at 29 time-points in a time course of 2 months in a longitudinal normality study. Since this is a normality study, the external disturbances of the environment of the monkeys are kept as low as possible. Therefore the variation in urine composition will mainly be caused by the natural variation in the metabolism of the monkeys. The objective of this research is obtaining a simplified view on the data, in which the metabolic biorhythms occurring in the chemical composition of the urine are captured.

The dataset consists of ^{1}H-NMR spectra of the described monkey urine. In addition to these spectra, information about the experimental error is present from repeated measurements on the urine samples. The experimental error is heteroscedastic, which means that the standard deviation of the experimental error depends on the size of the signal. A PCA model gives a distorted view on the data when the experimental error is non-uniform. WPCA is a method that can compensate for this non-uniform experimental error. The obtained WPCA model captures the natural variation underlying the data better than PCA (Wentzell, Andrews *et al.*, 1997). Therefore WPCA is used for the analysis of this dataset.

The error is described using a variance function (McCullaugh and Nelder, 1989). The weights used for the WPCA-analysis are calculated from the variance function. The results obtained from the analysis are compared to a data analysis with PCA, where the information about the error is not used.

In this paper the WPCA method is presented and its properties are compared to the properties of PCA. Then the application of WPCA to the analysis of the metabolomics dataset is shown. Finally the results of the WPCA analysis are compared to the results of the PCA analysis. The difference between the results obtained from PCA and WPCA is explained using the original data.

2.2 System and Methods

2.2.1 Urine Samples

Urine is obtained from rhesus monkeys (*Macaca mulatta*). Samples are taken of ten monkeys at 29 non-equidistant days over a time course of 57 days. Of the monkeys, 5 are male and 5 are female. Prior to NMR spectroscopic analysis the urine samples are lyophilised and pre-treated by adding 1mL of urine to 1 mL of sodium phosphate buffer (0.1 M, pH 6.0, made up with D_2O) containing 1mM sodium trimethylsilyl-[2,2,3,3,-$^2H_4$]-1-propionate (TSP) as an internal standard (δ_{TSP} = 0.0).

2.2.2 Data acquisition

NMR spectra are measured in triplicate using a Varian Unity 400 MHz NMR spectrometer using a proton-NMR setup at 293 K. Free Induction Decays (FIDs) are recorded as 64K datapoints with a spectral width of 8.000 Hz. 45 degree pulses are used with an acquisition time of 4.10 s and a relaxation delay of 2 s. The spectra are acquired by accumulation of 128 FIDs. The signal of the residual water is removed by a pre-saturation technique in which the water peak is irradiated with a constant frequency during 2 s prior to the acquisition pulse. The spectra are processed using the standard Varian software. An exponential window function with a line broadening of 0.5 Hz and a manual baseline correction is applied to all spectra. After referring to the internal NMR reference (TSP δ =0.0), peak shifts are corrected and line listings are prepared using WINLIN software (internal software TNO, see also (Vogels, Tas *et al.*, 1996)). Such preprocessing is a necessary step for subsequent data analysis. To obtain these listings all lines in the spectra above a threshold corresponding to about three times the signal-to-noise ratio are collected and converted to a data file suitable for multivariate data analysis applications. Each spectrum is normalised to have unit sum-of-squares, to remove differences in dilution between different urine samples.

NMR-spectra are obtained containing peaks on 332 chemical shifts from 0.89 to 9.13 ppm. Since three repeated measurements are performed on each sample, the total available data consists of 870 NMR-spectra. To avoid the problems connected with multiple sources of variation in the data each sample is represented by the mean spectrum of the repeated measurements (Jolliffe, 2002, pg. 351). In total there are 290 spectra in the dataset.

2.2.3 Data Analysis

2.2.3.1 Principal Component Analysis

The properties of PCA are well understood and thoroughly described in the literature (Jackson, 1991; Jolliffe, 2002). PCA defines a model of multi- or megavariate data. This model is a lower dimensional subspace that explains maximum variation in the original data. The dimension of this subspace is defined by the number of principal components that is chosen for the PCA-model. The loss function g that is minimised in PCA is given in equation (1).

$$(1) \quad g(\mathbf{T_{PCA}}, \mathbf{P_{PCA}} | \mathbf{X}) = \left\| \mathbf{X} - \mathbf{T_{PCA}} \mathbf{P_{PCA}^T} \right\|^2$$

where **X** is the $(I \times J)$ matrix containing the data, I is the number of samples (e.g. spectra) in the dataset and J is the number of variables (e.g. chemical shifts); $\mathbf{T_{PCA}}$ is the PCA score matrix of size $(I \times R)$ and $\mathbf{P_{PCA}}$ is the PCA loading matrix of size $(J \times R)$, where R is the number of principal components of the PCA-model. Each principal component is defined by the outer product of a column of $\mathbf{T_{PCA}}$ and the corresponding column of $\mathbf{P_{PCA}}$. The loading and the score matrices define the model obtained by PCA: The loadings-matrix is an orthonormal basis that describes the lower dimensional subspace of the PCA-model as a set of vectors in the space spanned by the variables. The scores describe each sample as a co-ordinate within the space spanned by the loadings. The lower-dimensional representation of the data in the PCA-model is easier to interpret than the original data and information about the phenomena underlying the variation in the data can be obtained from the PCA-model.

2.2.3.2 Weighted Principal Component Analysis

When there is additional information present about the data, PCA can not generally use this information. WPCA is a data analysis method that uses additional information about the data by the definition of weights. Using this information, a model containing scores and loadings is obtained that describes a space that is generally different from the space found with PCA. The weights in WPCA are introduced in the loss function. When there are no offsets in the data, the loss function h that is minimised in WPCA is given in equation (2) (Bro, Smilde *et al.*, 2002; Kiers, 1997; Wentzell, Andrews *et al.*, 1997).

(2) $$h\left(\mathbf{T_{WPCA}}, \mathbf{P_{WPCA}} \middle| \mathbf{X}, \mathbf{W}\right) = \left\| \mathbf{W} * \left(\mathbf{X} - \mathbf{T_{WPCA}}\mathbf{P_{WPCA}^{T}}\right)\right\|^{2}$$

The WPCA scores of dimensions $(I \times R)$ are denoted by $\mathbf{T_{WPCA}}$ and $\mathbf{P_{WPCA}}$ are the WPCA loadings of dimensions $(J \times R)$, where R is the number of principal components of the WPCA-model. The $*$ indicates the Hadamard (element-wise) product. Matrix **W** is the weight matrix containing weights corresponding to each element of the data and has dimensions $(I \times J)$. An element-wise expression of the minimisation function h is given in equation (3).

(3) $$h\left(\hat{x}_{ij} \middle| x_{ij}, w_{ij}\right) = \sum_{i=1}^{I}\sum_{j=1}^{J} w_{ij}{}^{2}\left(x_{ij} - \hat{x}_{ij}\right)^{2} = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(w_{ij}e_{ij}\right)^{2}$$

In equation (3) i from 1 to I is an index for the samples and j from 1 to J is an index for the variables. The value in the data for sample i and variable j is given by x_{ij}, the corresponding weight is given by w_{ij} and the estimated value of x_{ij} from the WPCA-model is given by $\hat{x}_{ij}$. The model residuals are given by e_{ij}.

Equation (3) shows that in WPCA, each value of e_{ij} receives its own weight w_{ij}. Comparison of equation (1) to equation (2) shows that PCA is a special case of WPCA, where all weights w_{ij} are equal to 1.

2.2.3.3 Properties of PCA and WPCA

The fact that in WPCA the model residuals are weighted means that WPCA is not equal to a scaling of each value x_{ij} in the data with a weight w_{ij}. Performing a PCA on the matrix product $\mathbf{W} * \mathbf{X}$ is not generally equal to solving the loss function given in equation (2). Only when the weight matrix **W** has a rank of 1, WPCA and performing a PCA on $\mathbf{W} * \mathbf{X}$ are equivalent (Bro and Smilde, 2003; Paatero and Tapper, 1993). Matrix **W** has a rank of 1 when WPCA is used as PCA or when scaling is applied to each variable or sample. When individual weights are defined for each x_{ij}, the rank of **W** is generally higher than 1. If the rank of **W** is higher than 1, a simple transformation of the variables or samples cannot solve equation (3) anymore.

A WPCA-model where the rank of **W** is higher than 1 is not nested (Wentzell, Andrews *et al.*, 1997). Therefore, the model $\mathbf{T_{WPCA}P_{WPCA}^T}$ with $R-1$ principal components is not contained within the model of R principal components. This means that the scores and loadings of every principal component will change when a different number of principal components is selected for the model. When WPCA-models of a different rank are compared, the information contained in all principal components should be considered simultaneously. Another difference between PCA and WPCA is the removal of the offsets in the data. Offsets are parts of the data that are constant for all samples. These offsets are often not interesting for the explanation of the variation in the data and a lower-rank model can be obtained by removing them. When PCA is used, offsets can be removed from the data by mean centering (Bro and Smilde, 2003; Gabriel, 1978; Kruskal, 1977). When using WPCA, mean-centering prior to the data analysis does not guarantee that the offsets in the data are removed. In WPCA the offsets have to be estimated along with the scores and loadings during the minimisation of the WPCA loss function.

2.2.3.4 Algorithms

Different algorithms are available that perform a data analysis using the loss function in equation (2) and provide a model of the obtained subspace containing orthogonal scores and orthonormal loadings. Gabriel and Zamir have developed an algorithm that performs criss-cross iterations (Gabriel and Zamir, 1979). Kiers has developed the PCAW algorithm (Kiers, 1997) that uses majorization iterations (Heiser, 1995) to minimise the loss function given in equation (2). Wentzell et al. have developed the Maximum Likelihood PCA (MLPCA) algorithm (Wentzell, Andrews *et al.*, 1997) that uses alternating least squares regression to perform WPCA. In MLPCA the weights can be defined to include covariances between the errors of the different values in the data. Bro et al. (Bro, Smilde *et al.*, 2002) have developed the MILES algorithm that combines the use of majorization iterations with the implementation of information about the covariances between the errors.

For this research, the algorithm developed by Kiers has been used, since it allows for an easy implementation of a general method for estimating the offsets in the data. This offset estimation is not implemented in the publicly available criss-cross and MLPCA algorithms. The method for offset estimation that is available in the MILES algorithm is restricted to estimation of column offsets and more general forms of offset estimation have not been implemented. Furthermore MILES requires the input of a $(IJ \times IJ)$ weight matrix which decreases the computational efficiency.

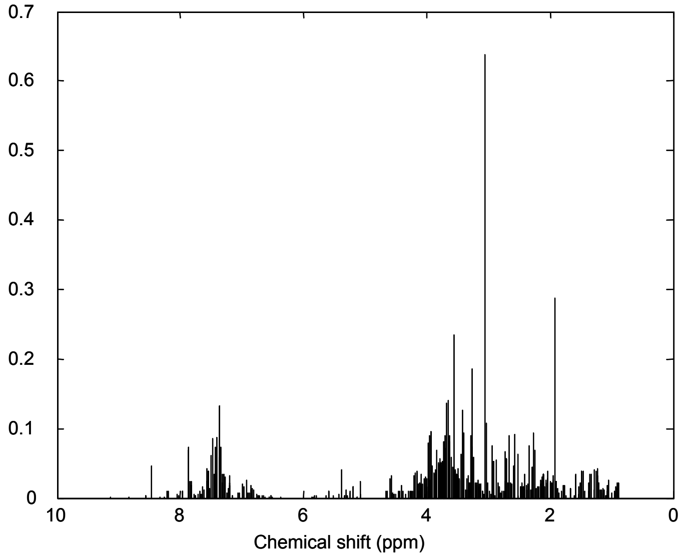

Figure 2 A typical 1H NMR-spectrum in the dataset

2.3 Implementation

For each monkey, a 29 x 332 data block was constructed. In these blocks each row contains an averaged spectrum of a urine sample collected at a time-point during the study. A typical spectrum in the dataset is given in Figure 2. The data blocks of all monkeys are collected to form a data matrix **X** containing 290 spectra on the rows. The variables in the matrix are the 332 chemical shifts in the NMR-spectra. The construction of the data matrix is given in Figure 3.

The model of the data should describe a low dimensional space, in which the normal longitudinal (time-dynamic) variation of the chemical composition of the monkey urine is expressed. This expression of the longitudinal variation in the data should be based on all monkeys in the dataset. The scores of a model of the data in **X** should contain longitudinal urine profiles for each monkey, expressed on the space of the normal longitudinal variation of all monkeys. The space spanned by the normal longitudinal variation of all monkeys is described by the loadings.

To make a model of the data that describes the longitudinal variation of the urine composition of all monkeys, the offsets belonging to each monkey need to be removed from the data. This can be achieved by column mean-centering each data block individually. Then the offset of each individual monkey will be removed and mean-centering of the data matrix **X** will be sustained (Bro and Smilde, 2003; Timmerman, 2001).

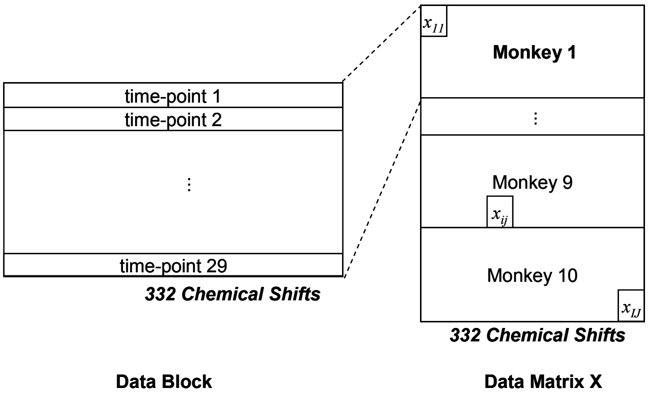

Figure 3 The data matrix X consists of 10 blocks, one for each monkey. Each block contains the spectra for all 29 time-points. X contains the data values x_{ij}

2.3.1 PCA-analysis

A PCA-model is made of the data matrix **X**. The rank of the PCA-model of matrix **X** can be determined by different methods. A simple and commonly used method is the 'scree graph' (Cattell, 1966;Jolliffe, 2002). The scree graph of this PCA model is given in Figure 4. Using this scree graph, 3 principal components are selected for the PCA-model. This PCA model explains 66 % of the variation in the data.

PCA is performed using the PCA-routine from the PLS-Toolbox 2.1 (Eigenvector Research, Inc., Manson, WA) for MATLAB (Mathworks, Inc., Natick, MA).

2.3.2 WPCA analysis

Prior to the WPCA data analysis, the elements w_{ij} of matrix **W** have to be determined. The weights in matrix **W** are calculated by using information from the repeated measurements of each sample. Matrix **W** is obtained by the analysis of the relationship between the size of the data values x_{ij} and the standard deviation of the repeated measurements used to calculate x_{ij}.

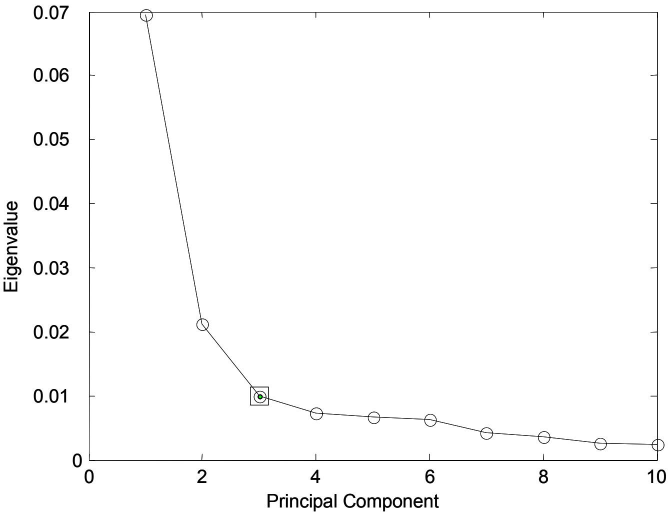

Figure 4 Scree graph of the PCA model. 3 principal components have been selected for the model from this graph

To make the estimation of the experimental error from the repeated measurements more robust, a binning procedure is performed on the data. Increasing values of x_{ij} are divided in 160 bins. The mean of x_{ij} and the mean standard deviation are a descriptor of each bin and their relationship is given in Figure 5. A first-order variance function is fitted through this relationship using the ROBUSTFIT function from the Statistics toolbox for MATLAB (Mathworks, Inc., Natick, MA). Figure 5 also shows this variance function.

The experimental error is estimated by calculating the variance function value for each x_{ij}. When the experimental error is small, the accuracy of x_{ij} is large and should be given a high importance in the data analysis. Therefore each w_{ij} is defined as the reciprocal of the estimate of the experimental error. The fitted function is independent of the number of predefined bins, since its parameters are stable when between 120 and 640 bins are used to estimate the function.

The insert in Figure 5 shows that for small signals with a peak size below 0.0030, the noise is relatively large (with a standard deviation up to 0.0035). Because we do not want to assign large weights to noise, the variance function is set to a constant value for signals lower than 0.0030. The signals below the threshold are not used to fit the variance function.

WPCA analysis is performed on the data matrix **X** where the data is not mean-centered, but where the offsets of each data block are estimated together with the WPCA-model. WPCA is performed using the PCAW algorithm by Kiers (Kiers, 1997). The algorithm has been adapted to include the estimation of the offsets of each data block within the algorithm. The used algorithm is given in the Appendix at the end of this chapter. The number of selected principal components for the WPCA-model is identical to that of the PCA-model to facilitate comparison of the model results. A WPCA model containing 3 principal components explains 57 % of the variation in the data. Because PCA and WPCA use different distance measures, this number cannot be directly compared to the amount of explained variation in PCA.

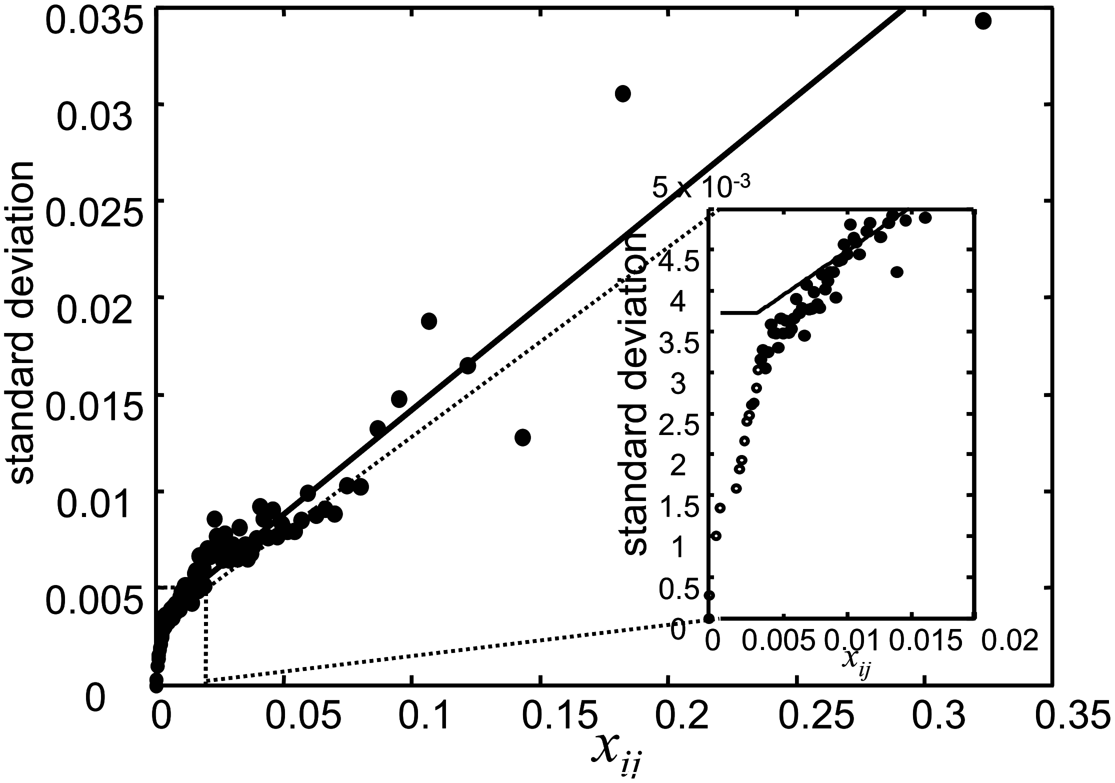

Figure 5 Relationship between the data value and its corresponding standard deviation. The continuous line indicates the estimated function and the closed circles indicate the points that have been used to estimate the function. The open circles indicate the points that have been discarded from the estimation of the function. The insert shows the lower threshold that is defined for the function.

2.3.3 Comparison of the PCA and the WPCA-model

The model obtained by PCA cannot be compared directly to the model obtained by WPCA. Both models describe a space that is spanned by a basis formed by the loadings.

The differences between the bases consist of differences in the subspaces spanned by the bases and the rotation of the two bases relative to each other. To compare the differences between both subspaces, the rotational differences between both bases need to be removed. Therefore an orthogonal transformation is performed on the WPCA-loadings to match the PCA-loadings as closely as possible. The transformation of the loadings is performed using a Procrustes rotation and reflection (Gower, 1995; Vandeginste, Massart *et al.*, 1998). The WPCA loadings are rotated to match the PCA loadings using equation (4).

(4) $\mathbf{P}^{\mathbf{T}}_{\mathbf{WPCA,Rot}} = \mathbf{Q}^{-1}\mathbf{P}^{\mathbf{T}}_{\mathbf{WPCA}}$

In equation (4), $\mathbf{P}_{\mathbf{WPCA,Rot}}$ are the rotated WPCA-loadings and $\mathbf{Q}$ is an $(R \times R)$ orthogonal transformation matrix. The same rotation has to be applied to the scores obtained from WPCA. After rotation both scores can be compared to each other. The scores are rotated using equation (5), where $\mathbf{T}_{\mathbf{WPCA,Rot}}$ are the rotated WPCA-scores.

(5) $\mathbf{T}_{\mathbf{WPCA,Rot}} = \mathbf{T}_{\mathbf{WPCA}}\mathbf{Q}$

2.4 Results

The effect of the weighting on the data analysis can be examined by investigating the difference between the obtained scores for PCA and WPCA.

From PCA and WPCA a model is obtained. Both models consist of a loading matrix containing 3 loading vectors. Also a score matrix is obtained that consists of 3 vectors of 290 elements. The 290 elements are the scores of the 29 time-points in the study for the 10 different monkeys. The scores obtained for PCA and WPCA are compared for male monkey 3, which is a typical monkey for the dataset. The PCA scores $\mathbf{T}_{\mathbf{PCA}}$ of monkey 3 are given in Figure 6 and Figure 7. The WPCA scores $\mathbf{T}_{\mathbf{WPCA,Rot}}$ of monkey 3 are given in Figure 8 and Figure 9.

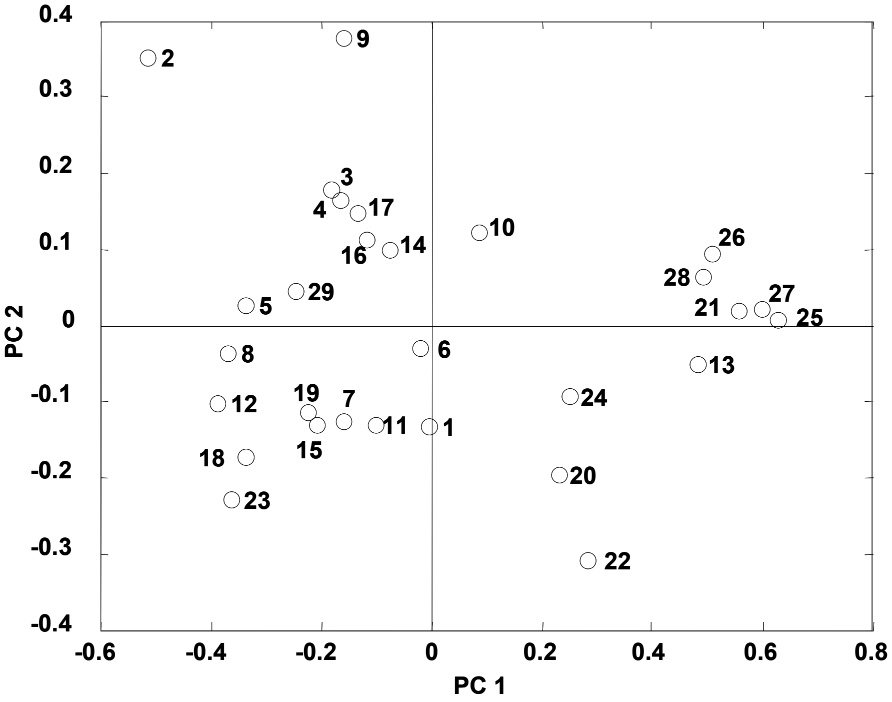

Figure 6 PCA scores for monkey 3 for PC 1 and PC 2

Comparing Figure 6 to Figure 8 shows that the scores of both methods are very similar for the first principal component. The difference between the scores in Figure 6 and Figure 8 is mostly in the scores on the second principal component. The second principal component of PCA is therefore different from the second principal component of WPCA.

Comparing Figure 7 to Figure 9 shows that there is a clear difference between the PCA and the WPCA models. For example, the scores on day 15 and day 19 are very similar in the PCA model, while the WPCA scores are quite different. Conversely, the WPCA scores of days 6 and 17 are very similar, while the PCA scores are quite different from each other.

The largest difference between both models for monkey 3 is the score of day 1. For time-point 1 the PCA-score has an extreme value for PC 3 and the WPCA score has an extreme value for PC 2. A possible explanation for the difference between PC 2 and PC 3 for time-point 1 is that the deviation of this spectrum from the mean spectrum of all days belonging to monkey 3 is very large for small and medium peaks, contrary to the large peaks. The scores of time-point 1 for monkey 3 are different for WPCA and PCA because smaller peaks with a smaller experimental error are given a higher importance in WPCA. Since the first principal component of WPCA and PCA are very similar, the loadings of PC 1 look very much alike for both methods. For PC 2 and PC 3 there are clear differences visible between the loadings. The loadings of PCA and WPCA for PC 3 are given in Figure 10, together with the standard deviation of each signal in the offset corrected data.

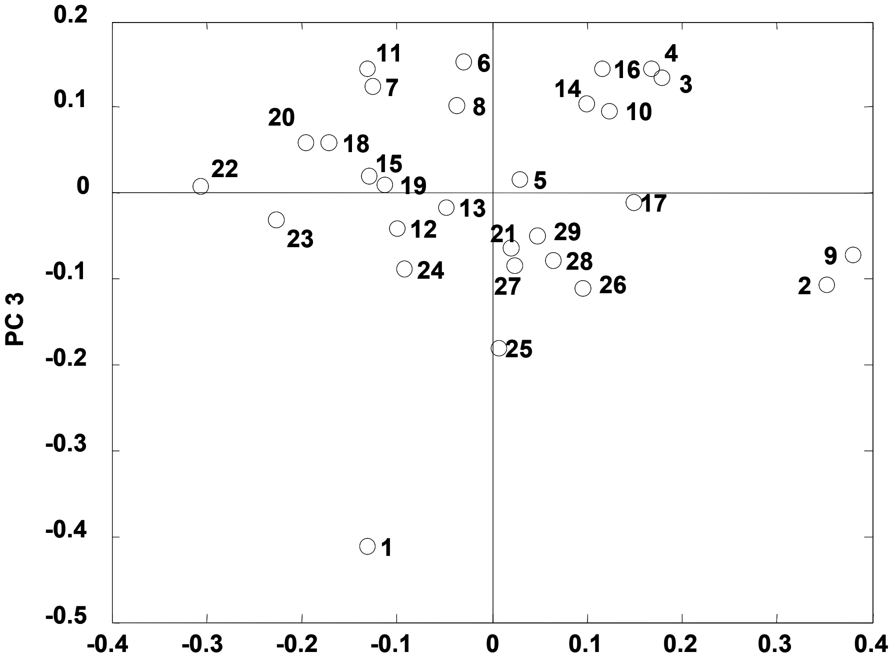

Figure 7 PCA scores for monkey 3 for PC 2 and PC 3

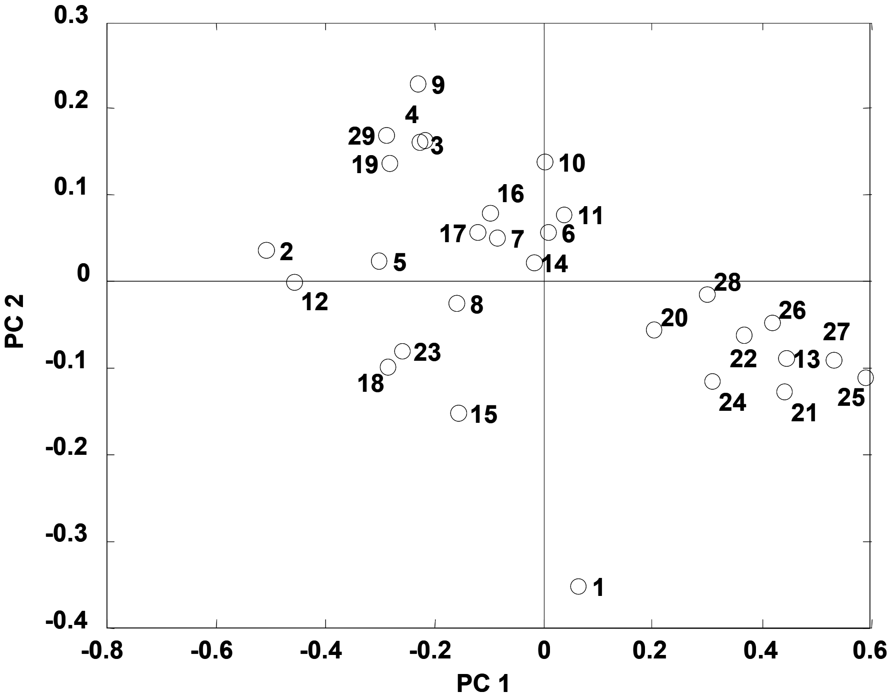

Figure 8 WPCA scores for monkey 3 for PC 1 and PC 2

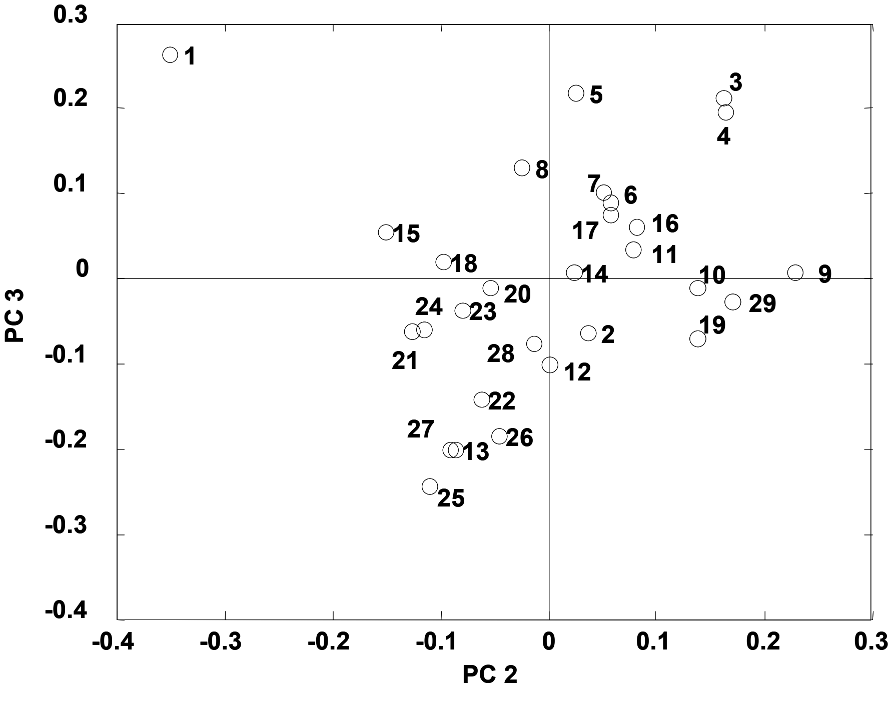

Figure 9 WPCA scores for monkey 3 for PC 2 and PC 3

It is clear from Figure 10 that in both models the chemical shifts on which there is a large natural variation are very important. The signals at e.g. chemical shifts 1.93, 2.93 and 3.56 have a large standard deviation and a large loading for both

PCA and WPCA. A clear difference between the loadings in Figure 10 is their direction: e.g. in the WPCA model the loading of 3.56 and 3.27 ppm are both positive, while in the PCA model the loading of 3.56 ppm is positive and the loading of 3.27 ppm is negative. Furthermore the large loading in PCA on 3.03 ppm is absent in the WPCA model. Finally, in the loadings of the PCA model only the large peaks are visible, while the loadings of the WPCA model give a higher importance to the region between 3.1 – 4.1 ppm. This also shows that WPCA gives a higher importance to the smaller peaks in the data that have a smaller experimental error. This is an example of natural variation that is obscured in the view obtained of the data by using PCA and is made visible in the WPCA model.

2.5 Conclusions

The scores and loadings show that the WPCA-model is different from the PCA-model. This is due to the fact that *a priori* information about the data is used in WPCA. The weighing used in this WPCA analysis is based on the experimental error: peaks that contain a smaller error are made more important in the data analysis. The model of the experimental error in Figure 5 shows that smaller peaks have a smaller experimental error. This means that smaller peaks are given a larger weight and therefore a higher importance in the data analysis. Additionally to accounting for the experimental error, a favourable side effect of weighing based on peak size is that it will decrease the bias of the model towards compounds with a high concentration in the urine. Comparison of the scores shows that the data analysis using WPCA gives a different view on the data than PCA. The WPCA model is more focused on the natural variation in the data.

The use of information about the experimental error in WPCA is one application where additional information about the data is used. Other possible applications of WPCA in metabolomics, proteomics and transcriptomics are abundant. Often prior information is present about the data. For example, data with many missing values can be obtained or information about the importance of certain variables can be present.

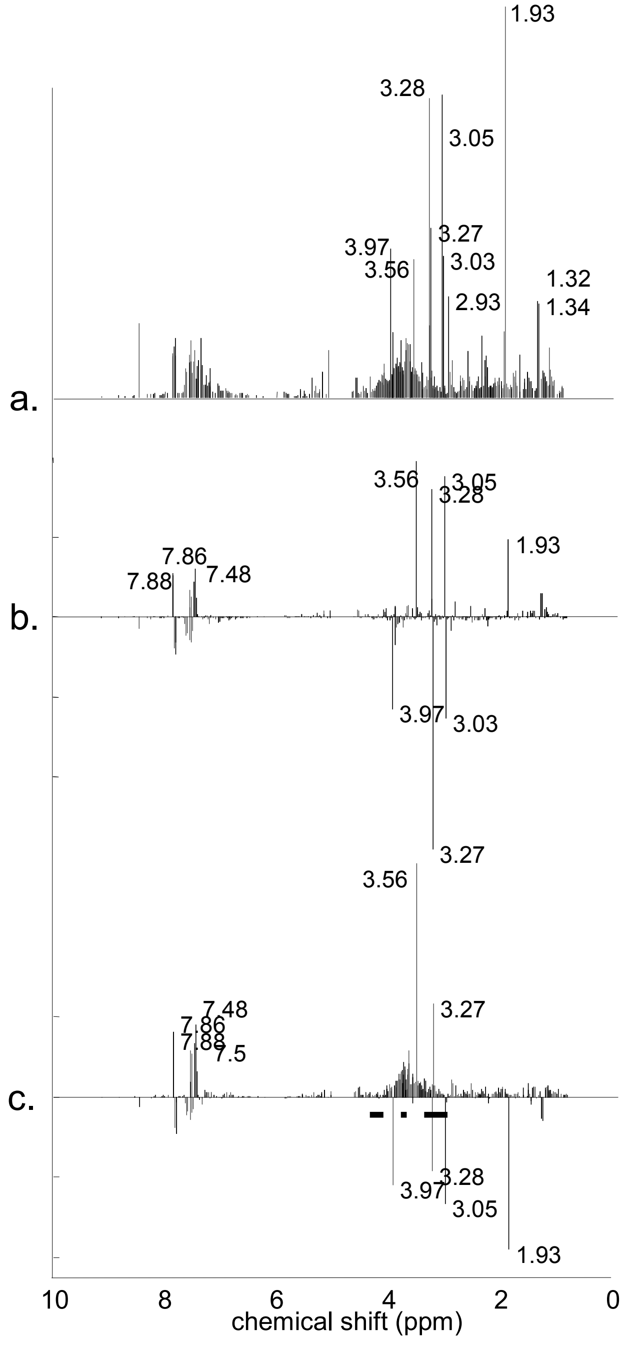

Figure 10 a. Standard deviation of the signal for each chemical shift in the data and loadings of PC 3 for b. PCA and c. WPCA. The horizontal dotted line in c. indicates the region between 3.1 and 4.1 ppm that has a higher importance in the WPCA model

Using additional information about the data obtained from an experiment, the researcher can generate weights for a WPCA analysis. WPCA can then be used as a generic bioinformatics tool that can use these sources of information in a data analysis.

2.6 Appendix

The adapted WPCA algorithm by Kiers is presented here. When the data contains column offsets, the algorithm minimises the function $\left\|\left(\mathbf{X}-\mathbf{1m}^{\mathbf{T}}-\mathbf{T}_{\mathbf{WPCA}}\mathbf{P}_{\mathbf{WPCA}}^{\mathbf{T}}\right)*\mathbf{W}\right\|^{2}$, where $\mathbf{m}^{\mathbf{T}}$ is a size J row-vector containing the offsets and **1** is a size I column vector containing ones. The minimisation is performed by alternatingly estimating $\mathbf{T}_{\mathbf{WPCA}}\mathbf{P}_{\mathbf{WPCA}}^{\mathbf{T}}$ from the offset corrected data $\mathbf{X}-\mathbf{1m}^{\mathbf{T}}$, and the offset vector $\mathbf{m}^{'}$ from the WPCA-model and the non-offset corrected data: $\mathbf{X}-\mathbf{T}_{\mathbf{WPCA}}\mathbf{P}_{\mathbf{WPCA}}^{\mathbf{T}}$. Both estimations are performed by majorization iterations (Heiser, 1995).

The algorithm is given by steps 1 to 11.

Initialisation

1. Initialise the algorithm using the PCA solution: $\mathbf{T}_{\mathbf{WPCA}}=\mathbf{T}_{\mathbf{PCA}}$, $\mathbf{P}_{\mathbf{WPCA}}=\mathbf{P}_{\mathbf{PCA}}$ and the column mean of **X**, $\mathbf{m}_{\mathbf{init}}^{\mathbf{T}}$ as the offset vector $\mathbf{m}^{\mathbf{T}}:\mathbf{m}^{\mathbf{T}}=\mathbf{m}_{\mathbf{init}}^{\mathbf{T}}$. Also a random initialisation can be chosen or an initialisation with all zeros for the scores and loadings.

2. Correct the data for the offset: $\mathbf{X}_{\mathbf{off}}=\mathbf{X}-\mathbf{1m}_{\mathbf{init}}^{\mathbf{T}}$, where $\mathbf{X}_{\mathbf{off}}$ is the offset corrected data and **1** is a size I column vector of ones.

Minimisation

3. Calculate the majorizing function **F** by: $\mathbf{F}=\dfrac{\mathbf{W}^{(2)}*\left(\mathbf{X}_{\mathbf{off}}-\mathbf{T}_{\mathbf{WPCA}}\mathbf{P}_{\mathbf{WPCA}}^{\mathbf{T}}\right)}{w_{\max}^{\mathbf{2}}}+\mathbf{T}_{\mathbf{WPCA}}\mathbf{P}_{\mathbf{WPCA}}^{\mathbf{T}}$ where $w_{\max}^{2}$ is the highest value in $\mathbf{W}^{(2)}$. And $\mathbf{W}^{(2)}=\mathbf{W}*\mathbf{W}$

4. Update $\mathbf{T}_{\mathbf{WPCA}}$ and $\mathbf{P}_{\mathbf{WPCA}}$ by performing a PCA on the matrix **F** calculated in 3: $\mathbf{F}=\mathbf{T}_{\mathbf{WPCA}}\mathbf{P}_{\mathbf{WPCA}}^{\mathbf{T}}+\mathbf{E}_{\mathbf{1}}$, where $\mathbf{E}_{\mathbf{1}}$ is a matrix containing the residuals.

5. Calculate $\mathbf{E}_{\mathbf{2}}$: a matrix containing the residuals and the offsets: $\mathbf{E}_{2}=\mathbf{X}-\mathbf{T}_{\mathbf{WPCA}}\mathbf{P}_{\mathbf{WPCA}}^{\mathbf{T}}$

6. Estimate the offsets by calculating the majorizing function $\mathbf{F_2}$: $\mathbf{F_2} = \frac{\mathbf{W^{(2)}} * \left(\mathbf{E}_2 - \mathbf{1m^T}\right)}{w_{\max}^2} + \mathbf{1m^T}$

7. Calculate the offset vector: $\mathbf{m^T} = \frac{1}{I}\mathbf{1^T F_2}$, where $\mathbf{1^T}$ is a size I row vector and $\mathbf{m^T}$ is the updated offset.

8. Update the offset corrected data: $\mathbf{X_{off}} = \mathbf{X} - \mathbf{1m^T}$

9. Evaluate the loss function $SS_{iter} = \left\|\mathbf{W} * \left(\mathbf{X_{off}} - \mathbf{T_{WPCA}P_{WPCA}^T}\right)\right\|^2$ where SS_{iter} is the value of the loss function at iteration $iter$

Evaluation

10. Calculate the convergence criterion: $dSS_{iter} = \frac{\left|SS_{iter} - SS_{iter-1}\right|}{SS_{iter-1}}$

11. Iterate steps 3 until 9 until convergence of dSS_{iter}

In our case, each block of the data contains a separate offset. This can be easily implemented in the presented algorithm.

2.7 References

Andrews, D. T. and Wentzell, P. D. (1997) Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer, *Anal.Chim.Acta*, 350, 341-352.

Bro, R. and Smilde, A. K. (2003) Centering and scaling in component analysis, *J.Chemometr.*, 17, 16-33.

Bro, R., Smilde, A. K., and Sidiropoulos, N. D. (2002) Maximum Likelihood fitting using ordinary least squares algorithms, *J.Chemometr.*, 16, 387-400.

R. B. Cattell (1966) The scree test for the number of components, *Multivar.Behav.Res.*, 1, 245-276

Gabriel, K. R. (1978) Least Squares Approximation of Matrices by Additive and Multiplicative Models, *J.R.Stat.Soc.B*, 40, 186-196.

Gabriel, K. R. and Zamir, S. (1979) Lower rank approximation of matrices by least squares with any choice of weights, *Technometrics*, 21, 489-498.

Glass, L. and Mackey, M. C., (1988), From Clocks to Chaos: "The Rythms of Life", Princeton University Press, Princeton

Gower, J. C. (1995) Orthogonal and Projection Procrustes Analysis, in Krzanowski, W. J. (eds), *Recent Advances in Descriptive Multivariate Analysis*, Oxford University Press, Oxford, pp. 113-134.

Heijne, W. H., Stierum, R. H., Slijper, M., Van Bladeren, P. J., and Van Ommen B. (2003) Toxicogenomics of bromobenzene hepatotoxicity: a combined transcriptomics and proteomics approach, *Biochem.Pharmacol.*, 65, 857-875.

Heiser, W. J. (1995) Convergent Computation by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis, in Krzanowski, W. J. (eds), *Recent Advances in Descriptive Multivariate Analysis*, Oxford University Press, Oxford, pp. 157-189.

Holmes, E. and Antti, H. (2002) Chemometric contributions to the evolution of metabonomics: mathematical solutions to characterising and interpreting complex biological NMR Spectra, *The Analyst*, 12, 1549-1557.

Holmes, E., Foxall, P. J. D., Nicholson, J. K., Neild, G. H., Brown, S. M., Beddell, C. R., Sweatman, B. C., Rahr, E., Lindon, J. C., Spraul, M., and Neidig, P. (1994) Automatic data reduction and pattern recognition methods for analysis of ^{1}H Nuclear Magnetic Resonance Spectra of Human Urine from Normal and Pathological states, *Anal.Biochem.*, 220, 284-296.

Jackson, J. E., (1991), A User's Guide to Principal Components, Wiley-Interscience, New York

Jolliffe, I. T., (2002), Principal Component Analysis, Springer-Verlag, New York

Kiers, H. A. L. (1997) Weighted least squares fitting using ordinary least squares algorithms, *Psychometrika*, 62, 251-266.

Kruskal, J. B. (1977) Some least squares theorems for matrices and N-way arrays, *Manuscript, Bell laboratories, Murray Hill, NJ*,

McCullaugh, P. and Nelder, J. A. (1989) The components of a generalized model, *Generalized Linear Models*, Chapman and Hall, London, New York, pp. 21-47.

Paatero, P. and Tapper, U. (1993) Analysis of different modes of factor analysis as least squares problems, *Chemometr.Intell.Lab.*, 18, 183-194.

Tas, A. C., van den Berg, H., Odink, J., Korthals, H., Thissen, J. T. N. M., and van der Greef, J. (1989) Direcht chemical ionization - mass spectrometric profiling in premenstrual syndrome, *J.Pharmaceut.Biomed.*, 7, 1239-1247.

Taylor, J., King, R. D., Altmann, T., and Fiehn, O. (2002) Application of metabolomics to plant genotype discrimination using statistics and machine learning, *Bioinformatics*, 18, S241-S248.

Timmerman, M. E., 2001, *Component Analysis of multisubject multivariate longitudinal data*, thesis, Rijksuniversiteit Groningen, pp. 70-71.

van der Greef, J., Davidov, E., Verheij, E. R., van der Heijden, R., Adourian, A. S., Oresic, M., Marple, E. W., and Naylor, S. (2003) The role of metabolomics in Systems Biology, in Harrigan, G. G. and Goodacre, R. (eds), *Metabolic Profiling: Its role in Biomarker Discovery and Gene Function Analysis*, Kluwer Academic Publishers, Boston/Dordrecht/London, pp. 170-198.

Vandeginste, B. G. M., Massart, D. L., Buydens, L. M. C., De Jong, S., Lewi, P. J., and Smeyers-Verbeke, J., (1998), Handbook of Chemometrics and Qualimetrics: Part B, Elsevier Science, Amsterdam

Vogels, J. T. W. E., Tas, A. C., Venekamp, J., and van der Greef, J. (1996) Partial linear fit: A new NMR spectroscopy preprocessing tool for pattern recognition applications, *J.Chemometr.*, 10, 425-438.

Wentzell, P. D., Andrews, D. T., Hamilton, D. C., Faber, K., and Kowalski, B. R. (1997) Maximum likelihood principal component analysis, *J.Chemometr.*, 11, 339-366.

3 Multilevel Component Analysis of Time-resolved Metabolic Fingerprinting data

3.1 Introduction

In genomics and systems biology a range of new methods has been developed that investigate the state of an organism or a cell on different aggregation levels. These methods form the basis for the analysis of the processes that lead from the DNA-code to phenotypic changes in an organism. Three of these methods are transcriptomics, proteomics and metabolic fingerprinting; all three generating large amounts of data. Contained within this data is the information about the organism that is investigated. To obtain this information, various types of data analysis methods are used.

Metabolic fingerprinting investigates the metabolism of an organism. Specifically, the effect of certain influences (e.g. diet, toxic stress or disease) on metabolism is the focus of research. In metabolic fingerprinting the chemical composition of cells, tissue or body fluids is studied. Within Life Sciences a major emphasis is on studying effects on the metabolite pattern in the development of diseases (biomarker or disease research) or the effect of drugs on this pattern, so called drug response profiling. The approach of metabolite fingerprinting originates from the nineteen-seventies and the combined approach of metabolite fingerprinting and multivariate statistics has been first applied in the early nineteen-eighties [1,2]. In the development of these strategies focusing on disease biomarker patterns, it became clear that biomarker patterns of living systems on a single time point or evaluated over different individuals generate important information but do not capture the dynamics of a system. Clear evidence has been generated that deregulation of the dynamics of a system can be the onset of disease development (dynamic disease concept, [3]). The key to investigate dynamic phenomena is the analysis of time series and a better understanding of "normality". Using non-invasive techniques like the analysis of urine (urinalysis) based on metabolic fingerprinting is an attractive approach and has been used in many studies since the early onset in clinical chemistry and toxicology related profiling. An example of the application of metabolic fingerprinting in a time-resolved experiment focusing on pre-menstrual syndrome (PMS) is given by Tas *et al.* [4].

A technique that is suitable for the analysis of the chemical composition of urine is ^{1}H-Nuclear Magnetic Resonance (NMR) spectroscopy [5]. Spectra of urine obtained by NMR-spectroscopy are complicated and have a high information density. Information about the phenomena underlying the variation in these spectra can be extracted using multivariate statistical methods commonly used in pattern recognition and chemometrics.

Much work has been done in both NMR spectroscopy and the multivariate statistical data analysis of urinalysis data in the field of sampling and instrumentation: e.g. [6,7,8,9,10,11] and in the development of models for the data analysis of metabolic fingerprinting data: e.g. [12,13,14,15].

The data generated in metabolic fingerprinting experiments on multiple individuals where time-resolved information is present is longitudinal, multisubject and multivariate. The data consists of multivariate spectra taken at multiple time-points in the study for several subjects (animals) simultaneously. Therefore, such datasets contain multiple types of variation: both originating from differences between the animals that are constant in time (e.g. due to differences in age, genotype) and the time-dynamic variation of the urine composition of each individual animal (e.g. biorhythms, onset of disease or a toxic insult). This means these datasets have a multilevel structure [16].

Research where data with a multilevel structure is generated is abundant in biology, as well as in many other fields. Examples of multilevel problems are the monitoring of hospital patients in time, the monitoring of batch processes in process industry and economic time-series analysis of multiple countries, economic branches or companies.

Principal Component Analysis (PCA) is a method that is often used for the unsupervised analysis of metabolic fingerprinting data. It gives a simplified lower-dimensional representation of the variation that is present in a dataset. The scores and loadings obtained from PCA can be visualized and interpreted.

However, if PCA is used for the analysis of multilevel data, the different types of variation in the multilevel data will not be separated and the obtained principal components will describe a mixture of different types of variation. In the analysis of time-resolved metabolic fingerprinting data this means that a PCA model does not give a separate interpretation of the time-dynamic variation of the urine composition of the individuals and the variation between the individuals. Both types of variation in the data are confounded within the PCA model, which seriously hampers the interpretation of the phenomena underlying the variation in the data.

Kiers *et al.* [17] have proposed a generalization of PCA to analyze data containing sets of samples belonging to multiple populations, measured on the same variables. This method is called Simultaneous Component Analysis (SCA). Prior to SCA, the static variation between the populations is removed from the data. Therefore the SCA model is focused on the variation within the populations and this variation is not confounded with the variation between populations, like in the PCA model. Timmerman *et al.* have applied this method to the analysis of multivariate multisubject time-resolved data [18]. An SCA model gives a better view on the time-resolved variation in these datasets. However if SCA is used, interesting information about the static variation between individuals is lost from the analysis. Therefore possibly valuable information about the processes underlying the variation in the data is not obtained using the SCA model.

Multilevel methods make a model of a dataset that contains different submodels for the different types of variation in the data. Timmerman has proposed a method called Multilevel Component Analysis (MCA) [19], in which various component submodels give a summary of the different types of variation in the data. A constrained version of the MCA model is the Multilevel *Simultaneous* Component Analysis (MSCA) model. In this model the SCA and MCA approaches are combined. As a result a two-level MSCA model of time-resolved measurements on multiple individuals consists of two submodels: An SCA model describing the dynamic variation of the individuals (the within-individual variation) and a PCA model describing the static differences between the individuals (the between-individual variation).

The use of MSCA for the analysis of metabolic fingerprinting data is illustrated by the following application. The urine of healthy rhesus monkeys has been sampled at 29 time-points in a time course of 2 months in a normality study. Since this is a normality study, the external disturbances of the environment of the monkeys are kept as low as possible. The variation of the metabolism of the monkeys is monitored by ^{1}H-NMR-spectroscopy of urine of the monkeys. These spectra are analyzed using MSCA. This model is used for investigating biorhythms in the urine composition and therefore in the metabolism of the monkeys. The information about these biorhythms can then be used to determine whether variation that is observed in e.g. a toxicology study can be attributed to the toxic insult or is part of the normal variation in metabolism. Also the phenomena underlying the normal biological variation in the urine composition between different monkeys can be identified using this MSCA model.

The remainder of the article is organized as follows. First PCA is described briefly. Then the MCA method is explained and the constraints on the MCA model are defined that lead to the formulation of the MSCA model. Subsequently, MSCA is used for the analysis of the monkey urine data and the results of the data analysis are explained. Finally, the results obtained from MSCA are compared to the results of an analysis using PCA. In subsequent chapters MSCA is generalized for the use with any ANOVA model (MSCA builds on the one-way ANOVA model); this more general merger of ANOVA and SCA will be referred to as ANOVA-SCA, or ASCA.

3.2 Materials and Methods

3.2.1 Urine Samples

Urine is obtained from rhesus monkeys (*Macaca mulatta*). Samples are taken of ten monkeys at 29 non-equidistant days over a time course of 57 days. Of the monkeys, 5 are male and 5 are female. Prior to NMR spectroscopic analysis the urine samples are lyophilized and pre-treated by adding 1mL of urine to 1 mL of sodium phosphate buffer (0.1 M, pH 6.0, made up with D_2O) containing 1mM sodium trimethylsilyl-[2,2,3,3,-$^2H_4$]-1-propionate (TSP) as an internal standard (δ_{TSP} = 0.0).

3.2.2 Data acquisition

NMR spectra are measured in triplicate using a Varian Unity 400 MHz NMR spectrometer using a standard proton-NMR setup at 293 K. Free Induction Decays (FIDs) are recorded as 64K datapoints with a spectral width of 8.000 Hz. A single 45 degree pulse is used with an acquisition time of 4.10 s and a relaxation delay of 2 s. The spectra are acquired by accumulation of 128 FIDs. The signal of the residual water is removed by a pre-saturation technique in which the water peak is irradiated with a constant frequency during 2 s prior to the acquisition pulse. The spectra are processed using the standard Varian software. An exponential window function with a line broadening of 0.5 Hz and a manual baseline correction is applied to all spectra. After referring to the internal NMR reference TSP. The NMR data reduction file was imported into Winlin (V1.10, TNO, The Netherlands). Minor variations from comparable signals in different NMR spectra were adjusted and lines were fitted without loss of resolution, based on the Partial Linear Fit Algorithm [20]. Subsequently, all lines in the spectra above a threshold corresponding to about three times the signal-to-noise ratio are collected and converted to a data file suitable for multivariate data analysis applications. Each spectrum is vector normalized.

NMR-spectra are obtained containing peaks on 332 chemical shifts from 0.89 to 9.13 ppm. Since three repeated measurements are performed on each sample, the total available data consists of 870 NMR-spectra. To avoid the problems connected with multiple sources of variation in the data, each sample is represented by the mean spectrum of the repeated measurements [21, pg. 351]. In total there are 290 spectra in the dataset.

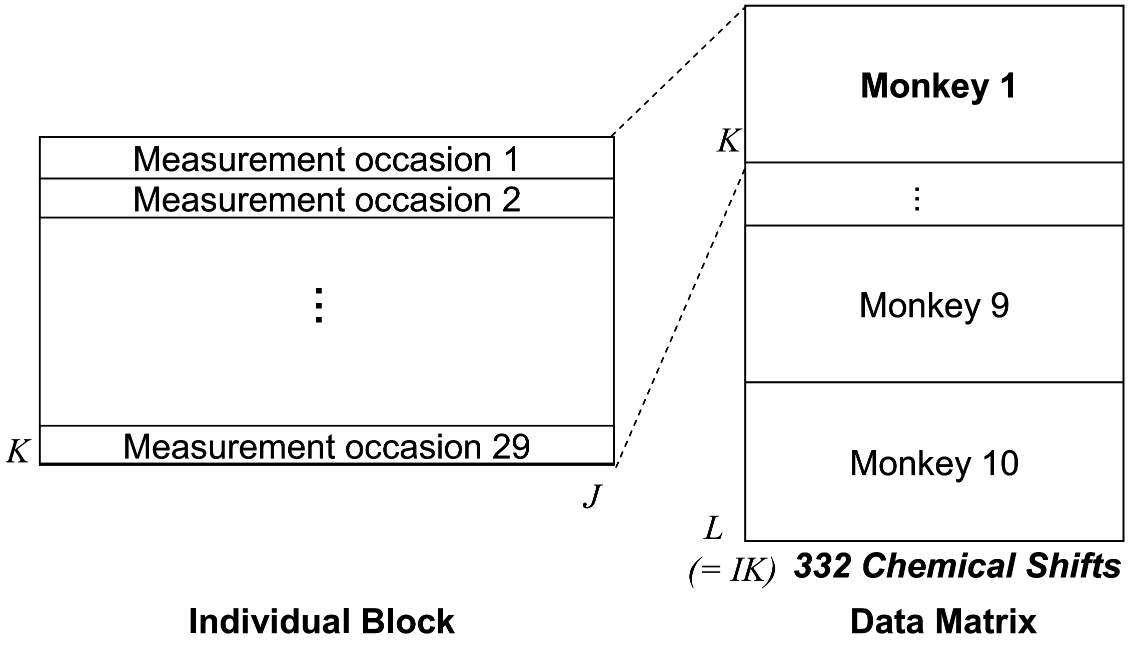

Figure 11 Structure of the dataset. K is the total number of measurement occasions in the study (that is equal for all monkeys). I is the total number of individuals (monkeys) and L is the total number of samples (spectra) in the dataset. J is the total number of variables (chemical shifts) in the data.

3.2.3 Data Analysis

A matrix $\mathbf{X}_{\text{raw}}$ of dimensions $(L \times J)$ is constructed that contains the obtained NMR-spectra, where L is the number of samples (e.g. spectra) in the dataset and J is the number of variables (e.g. chemical shifts). For the notation used here, see the Appendix. The structure of $\mathbf{X}_{\text{raw}}$ is given in Figure 11. The multivariate data analysis methods that are described here, all attempt to approximate the information that is contained in a data matrix $\mathbf{X}_{\text{raw}}$ by defining a model that contains one or more lower-dimensional subspaces within the high-dimensional data. This model should focus on the variation in the data. Therefore the invariant part of the data (the offset) is removed by mean-centering: thereby each column mean of matrix $\mathbf{X}_{\text{raw}}$ is modeled separately from the column variation [22,23].

The analysis methods used in this research attempt to approximate the information in $\mathbf{X}_{\mathbf{raw}}$, by fitting component models that estimate this data. This estimation is given in equation (1).

(1) $\mathbf{X}_{\mathbf{raw}} = \mathbf{1m}^{\mathbf{T}} + \hat{\mathbf{X}} + \mathbf{E}$

where $\mathbf{m}^{\mathbf{T}}$ is a size J row vector containing the mean values of each column of $\mathbf{X}_{\mathbf{raw}}$ and $\mathbf{1}_L$ is a size L column vector containing ones (the term $1_L\mathbf{m}^{\mathbf{T}}$ is related to the mean centering); matrix $\hat{\mathbf{X}}$ of dimensions $(L \times J)$ contains the data as it is estimated by the model. In equation (1), the parts of the data (e.g. small sources of variation, noise) that are not contained in the model are given in de model residuals $\mathbf{E}$ of dimensions $(L \times J)$. Fitting the models boils down to minimizing the sum of squares $\|\mathbf{E}\|^2$ [24].

3.2.3.1 Principal Component Analysis

The properties of PCA are well understood and thoroughly described in the literature [21,25]. PCA generates a bilinear model of multivariate data, by defining a lower dimensional subspace that explains as much variation in the original data as possible. The number of principal components that is chosen for the PCA model defines the dimension of this subspace. The model that is constructed in PCA is given in equation (2).

(2) $\mathbf{X}_{\mathbf{raw}} = \mathbf{1}_L\mathbf{m}^{\mathbf{T}} + \mathbf{T}_{\mathbf{PCA}}\mathbf{P}_{\mathbf{PCA}}^{\mathbf{T}} + \mathbf{E}_{\mathbf{PCA}}$

where $\mathbf{T}_{\mathbf{PCA}}$ is the PCA score matrix of size $(L \times R)$ and $\mathbf{P}_{\mathbf{PCA}}$ is the PCA loading matrix of size $(J \times R)$, where R is the number of principal components of the PCA model. The matrix $\mathbf{E}_{\mathbf{PCA}}$ of size $(L \times J)$ contains the residuals of the PCA model.

Each principal component is given by the outer-product of a column of $\mathbf{T}_{\mathbf{PCA}}$ and the corresponding column of $\mathbf{P}_{\mathbf{PCA}}$. The loading and the score matrices define the model obtained by PCA. The columns of the loading matrix form an orthonormal basis that describes the lower dimensional subspace of the PCA model as a set of vectors in the space spanned by the variables. The scores describe each sample as a co-ordinate within the space spanned by the loadings. The lower-dimensional representation of the data in the PCA model is easier to interpret than the original data and information about the phenomena underlying the variation in the data can be more easily obtained from the PCA model. Equation (2) shows that a single lower-dimensional subspace is defined in the model to describe all variation in the data.

3.2.3.2 Multilevel (Simultaneous) Component Analysis

Multilevel Component Analysis (MCA) is an extension of PCA that can be used if the variation in the data occurs on different levels simultaneously, e.g. on two levels when a model is made of the time-dynamic variation of different individuals, like the urine composition of different monkeys on multiple time-points in this study. A two-level MCA can also be defined as a model within the framework of equation (1), where the data matrix $\mathbf{X}_{\mathbf{raw}}$ is reconstructed by an offset and models describing the variation on 2 levels. The equation for the MCA model for one individual i (where $i = 1 \ldots I$) is given in equation (3).

$$(3) \quad \mathbf{X}_{\mathbf{raw},i} = \mathbf{1}_{K_i}\mathbf{m}^{\mathbf{T}} + \mathbf{1}_{K_i}\mathbf{t}_{\mathbf{b},i}^{\mathbf{T}}\mathbf{P}_{\mathbf{b}}^{\mathbf{T}} + \mathbf{T}_{\mathbf{w},i}\mathbf{P}_{\mathbf{w},i}^{\mathbf{T}} + \mathbf{E}_{\mathbf{MCA},i}$$

where $\mathbf{X}_{\mathbf{raw},i}$ is the $(K_i \times J)$ part of $\mathbf{X}_{\mathbf{raw}}$ that belongs to individual i; $\mathbf{t}_{\mathbf{b},i}^{\mathbf{T}}$ is a size $R_{\mathbf{b}}$ row vector containing the between-individual scores of individual i; $\mathbf{P}_{\mathbf{b}}$ is a $(J \times R_{\mathbf{b}})$ matrix containing the between-individual loadings and $R_{\mathbf{b}}$ is the number of components chosen for the between-individual model. The within-individual scores of i are denoted by $\mathbf{T}_{\mathbf{w},i}$, which is a $(K_i \times R_{\mathbf{w},i})$ matrix; $\mathbf{P}_{\mathbf{w},i}$ is a $(J \times R_{\mathbf{w},i})$ matrix containing the orthogonal within-individual loadings of individual i and $R_{\mathbf{w},i}$ is the number of components chosen for the within-individual model of individual i (this means each within-individual model can have a different number of components); $\mathbf{1}_{K_i}$ denotes a size K_i column vector of ones and K_i is the number of measurement time-points (occasions) for individual i. The matrices $\mathbf{T}_{\mathbf{w},i}$ are constrained to be mean-centered: $\mathbf{1}_{K_i}^{\mathbf{T}}\mathbf{T}_{\mathbf{w},i} = \mathbf{0}_{R_{\mathbf{w},i}}^{\mathbf{T}}$. Therefore all within-individual models are orthogonal to the between-individual model. A mathematical proof of this has been given by Timmerman [19]. The constraint given to the between-individual scores is that $\sum_{i=1}^{I} K_i \mathbf{t}_{\mathbf{b},i} = \mathbf{0}_{R_\mathbf{b}}$. By imposing this constraint the between-individual scores will describe the deviation of each individual from the overall mean $\mathbf{m}^{\mathbf{T}}$.

Equation (3) shows that the data in matrix $\mathbf{X}_{\mathbf{raw}}$ is reconstructed in the MCA model by an offset term $\mathbf{1}_L\mathbf{m}^{\mathbf{T}}$ that is equal for all samples in the data, a between-individual part $\mathbf{1}_{K_i}\mathbf{t}_{\mathbf{b},i}^{\mathbf{T}}\mathbf{P}_{\mathbf{b}}^{\mathbf{T}}$ that is equal for all samples belonging to an individual i (hence the vector $\mathbf{1}_{K_i}$) and a within-individual part $\mathbf{T}_{\mathbf{w},i}\mathbf{P}_{\mathbf{w},i}^{\mathbf{T}}$ that is different for each sample in the data. This is a familiar concept from ANOVA [26]. Hence, MCA can be seen as a ‘merger’ of the factor estimation aspect of ANOVA and PCA.

The I loading matrices $\mathbf{P}_{\mathbf{w},i}$ of the within-individual models are generally different from each other. This implies that the within-individual scores $\mathbf{T}_{\mathbf{w},i}$ of the different individuals cannot be compared directly between individuals. This seriously hampers the interpretability of the MCA model. To increase the interpretability of such a multilevel component model the within-individual variation of all individuals is expressed on the same basis. When the within-individual loadings are imposed to be equal for all individuals, the MSCA model is obtained. As a result, the two-level MSCA model only contains two different subspaces in which different types of variation in the data are explained. In the MSCA model, the within-individual scores of different individuals can be directly compared to each other, since they are expressed on the same basis. The MSCA model is given in equation (4).

(4) $\mathbf{X}_{\mathbf{raw},i} = \mathbf{1}_{K_i}\mathbf{m}^{\mathbf{T}} + \mathbf{1}_{K_i}\mathbf{t}_{\mathbf{b},i}^{\mathbf{T}}\mathbf{P}_{\mathbf{b}}^{\mathbf{T}} + \mathbf{T}_{\mathbf{w},i}\mathbf{P}_{\mathbf{w}}^{\mathbf{T}} + \mathbf{E}_{\mathbf{MSCA},i}$

where $\mathbf{P}_{\mathbf{w}}$ is a $(J \times R_{\mathbf{w}})$ matrix containing the within-individual loadings that are equal for all individuals. This means that each within-individual MSCA model contains an equal number of components R_w. For the model in equation (4) the MCA constraints hold as well: $\sum_{i=1}^{I} K_i \mathbf{t}_{\mathbf{b},i}^{\mathbf{T}} = \mathbf{0}_{R_{\mathbf{b}}}$ and $\mathbf{1}_{K_i}^{\mathbf{T}}\mathbf{T}_{\mathbf{w},i} = \mathbf{0}_{R_{\mathbf{w}}}^{\mathbf{T}}$. Note that for simplicity both in equations (3) and (4), $\mathbf{T}_{\mathbf{w},i}$ is used to indicate the within-individual scores of individual i. Although the symbols are identical, the values of $\mathbf{T}_{\mathbf{w},i}$ may differ between both methods.

In the MSCA model in equation (4), the non-dynamic variation between the individuals is explained by the term $\mathbf{1}_{K_i}\mathbf{t}_{\mathbf{b},i}^{\mathbf{T}}\mathbf{P}_{\mathbf{b}}^{\mathbf{T}}$ (that differs between individuals) and the dynamic variation of each individual is explained by the term $\mathbf{T}_{\mathbf{w},i}\mathbf{P}_{\mathbf{w}}^{\mathbf{T}}$. This means that in the MSCA model, the static and dynamic variations are separately described and are not confounded. When more than two types of variation are of interest to the experiment, the MSCA model can be extended to include more than two levels. Then each level will separately describe these types of variation.

In general the subspace spanned by the PCA loadings differs from the subspaces spanned by the loadings of the MSCA submodels: therefore generally the MSCA model cannot be obtained by any rotation of the PCA model.

The model described in equation (4) is the basis for a range of methods. These methods differ in the additional constraints that are defined for covariances of the within-individual scores $\mathbf{T}_{\mathbf{w},i}$. The most general MSCA model is the MSCA-P model. In this model no constraints are placed on the within-individual scores, apart from the mean centering of the scores. The MSCA-P model determines the directions in which the within-individual variation is largest for *all* individuals and takes these directions as the within-individual loadings. This means that for each individual, the correlation between the within-individual components can differ. Also the variation described by each component for each individual can be different. Therefore the MSCA-P model does not make any assumptions about similarities between the time-dynamic variations of different individuals.

Timmerman has given several additional constraints for the MSCA-P model. The MSCA-PF2, IND and ECP models are all increasingly constrained to incorporate assumptions about the relationships between the within-individual variation of different individuals. The mathematical basis for these MSCA models has been thoroughly described by Timmerman and Timmerman *et al*. [19,18].

For all MSCA models using the appropriate constraints, the parameters can be obtained by a least-squares method. This means that the MSCA model can be obtained from the minimization of the criterion $f_{\mathbf{MSCA}}$ given in equation (5), under all constraints defined for the MSCA model.

$$(5)\qquad f_{\mathbf{MSCA}}\left(\mathbf{m}^{\mathbf{T}},\mathbf{t}_{\mathbf{b},i},\mathbf{P}_{\mathbf{b}},\mathbf{T}_{\mathbf{w},i},\mathbf{P}_{\mathbf{w}}\right)=\sum_{i=1}^{I}\left\|\mathbf{X}_{\mathbf{raw},i}-\mathbf{1}_{K_i}\mathbf{m}^{\mathbf{T}}-\mathbf{1}_{K_i}\mathbf{t}_{\mathbf{b},i}^{\mathbf{T}}\mathbf{P}_{\mathbf{b}}^{\mathbf{T}}-\mathbf{T}_{\mathbf{w},i}\mathbf{P}_{\mathbf{w}}^{\mathbf{T}}\right\|^2$$

Comparison of equation (4) (which has to be concatenated for all individuals) to equation (2) shows that all MSCA methods are constrained versions of the PCA model. Hence, the PCA model will explain equal or more variation in the data than a MSCA model in which an equal number of components (cumulative for all submodels) is fitted. The more constrained the MSCA model is, the less variation in the data it will explain for an equal number of fitted components. However, the MSCA models have a much better interpretability for multilevel data than the PCA model due to the fact that the different submodels of MSCA describe different types of variation in the data. The loss of fit can therefore be seen as a payoff for an increased interpretability of the model. Comparison of equation (4) to equation (2) shows furthermore that the PCA and the MSCA-P models are equal when there is no between-individual variation in the data.

Due to the orthogonality of the submodels, the criterion $f_{\mathbf{MSCA}}$ can be minimized by separately determining the within and between-individual models [19]. How to obtain the MSCA model parameters in the case of an equal number of measurement occasions for all individuals is shown in the next paragraph.

3.2.3.3 Obtaining the MSCA model parameters

In the remainder of the article an equal number of measurement occasions $K_1 = K_2 = K_i = K$ is assumed for each individual, so that $IK = L$.

The between-individual model and the within-individual model can be determined after a decomposition of matrix $\mathbf{X}$, where matrix $\mathbf{X}$ contains the mean-centered data and has dimensions $(L \times J)$. This decomposition consists of a mean-centering of each matrix $\mathbf{X}_i$, where $\mathbf{X}_i$ is the $(K \times J)$ partition of matrix $\mathbf{X}$ belonging to individual i. This is given in equation (6).

$$(6) \quad \mathbf{X}_{\mathbf{c},i} = \mathbf{X}_i - \mathbf{1}_K \mathbf{m}_i^{\mathbf{T}}$$

where $\mathbf{m}_i^{\mathbf{T}}$ is a row vector containing the column means of matrix $\mathbf{X}_i$ and $\mathbf{X}_{\mathbf{c},i}$ is the size$(K \times J)$ matrix containing only the dynamic within-individual variation belonging to individual i.

The I vectors $\mathbf{m}_i^{\mathbf{T}}$ can be concatenated into a $(I \times J)$ matrix $\mathbf{M}$. Matrix $\mathbf{M}$ now contains the non-dynamic differences between the individuals: the between-individual variation. In the case of an equal number of measurement occasions for each individual, the between-individual model can be determined by performing a PCA on $\mathbf{M}$. This is given in equation (7). The parameters of the between-individual model for different numbers of measurement occasions for each individual can be calculated by the method given by Timmerman [19].

$$(7) \quad \mathbf{M} = \mathbf{T_b} \mathbf{P_b^T} + \mathbf{E_b}$$

where $\mathbf{T_b}$ is a $(I \times R_{\mathbf{b}})$ matrix of which the rows contain the between-individual scores $\mathbf{t}_{\mathbf{b},i}^{\mathbf{T}}$ and $\mathbf{E_b}$ is an $(I \times J)$ matrix containing the residuals of the between-individual model. Since matrix $\mathbf{X}$ is mean-centered, matrix $\mathbf{M}$ is also mean-centered and therefore the between-individual scores $\mathbf{T_b}$ are mean-centered as well. Hence the constraint $\sum_{i=1}^{I} K_i \mathbf{t}_{\mathbf{b},i} = \mathbf{0}_{R_\mathbf{b}}$ is met.

The within-individual submodel can be calculated by fitting a PCA (or a PF2, IND or ECP model [18]) on $\mathbf{X_c}$. This is given in equation (8), where the desired MSCA constraints are defined on $\mathbf{T}_{\mathbf{w},i}$.

$$(8) \quad \mathbf{X_c} = \mathbf{T_w}\mathbf{P_w^T} + \mathbf{E_w}$$

where $\mathbf{X_c}$ is a $(L \times J)$ matrix in which all matrices $\mathbf{X}_{\mathbf{c},i}$ are concatenated, $\mathbf{T_w}$ is a $(L \times R_\mathbf{w})$ matrix in which the within-individual scores $\mathbf{T}_{\mathbf{w},i}$ are concatenated and $\mathbf{E_w}$ is a $(L \times J)$ matrix containing the residuals of the within-individual model. Since all matrices $\mathbf{X}_{\mathbf{c},i}$ are centered, matrix $\mathbf{X_c}$ is centered. Therefore the score matrices $\mathbf{T}_{\mathbf{w},i}$ (and therefore $\mathbf{T_w}$) are centered by construction.

In MSCA-P the least squares criterion $f_{\mathbf{MSCA}}$ can be minimized by calculating the model $\mathbf{T_w}\mathbf{P_w^T}$ performing a PCA on $\mathbf{X_c}$: therefore in the case of an equal number of measurement occasions for all individuals, the MSCA-P model is a combination of two PCA models and describes as much variation as possible both on the between and the within-individual level. The more constrained MSCA-PF2, IND and ECP models cannot be expressed as a combination of PCA submodels.

3.2.3.4 Explained variation in MSCA

The decomposition of $\mathbf{X}$ into two orthogonal parts, given in equation (6) for a single individual, shows that the variation in $\mathbf{X}$ can be separated into a part of between-individual variation and a part of within-individual variation, similar to ANOVA [26]. This is given in equation (9).

(9) $\left\|\mathbf{X}\right\|^2 = \left\|\mathbf{X_c}\right\|^2 + K \times \left\|\mathbf{M}\right\|^2$

The sums-of-squares $\left\|\mathbf{X_c}\right\|^2$ and $K \times \left\|\mathbf{M}\right\|^2$ can be used to determine the magnitudes of the within and between-individual variation.

Like in PCA, the percentage of explained variation can be used as a measure of fit of the model to the data. Because the two-level MSCA model consists of two independent submodels, three different criteria of fit can be defined: the percentage of explained variation of the within-individual submodel, of the between-individual submodel and of the entire MSCA model. Equations (7) and (8) show that the within-individual model spans a subspace of $\mathbf{X_c}$ and the between-individual model spans a subspace of $\mathbf{M}$. Therefore the percentage of explained variation for each MSCA submodel can be calculated analogously to PCA. Furthermore, calculating the percentage of explained within-individual variation of each individual can identify individuals that are not well described by the within-individual level model.

3.3 Results and discussion

3.3.1 Model selection

Analysis of the variation in $\mathbf{X}$ by using equation (9) shows that 24 % is between-individual information and 76 % is within-individual variation. Both the between-individual variation and the within-individual variation have a considerable magnitude and therefore a Multilevel Component Analysis method should be used for the analysis of the monkey urine data to obtain a view on the data in which both types of variation are not confounded. To increase the interpretability of the model, the time-resolved variation of the urine composition of the different monkeys can be expressed in the same subspace by selecting a two-level MSCA model for the analysis of this data.

In the analysis of the monkey urine data, no assumptions are made about the relationship between the within-individual variation of different individuals. Therefore a two-level MSCA-P model is selected for the data analysis.

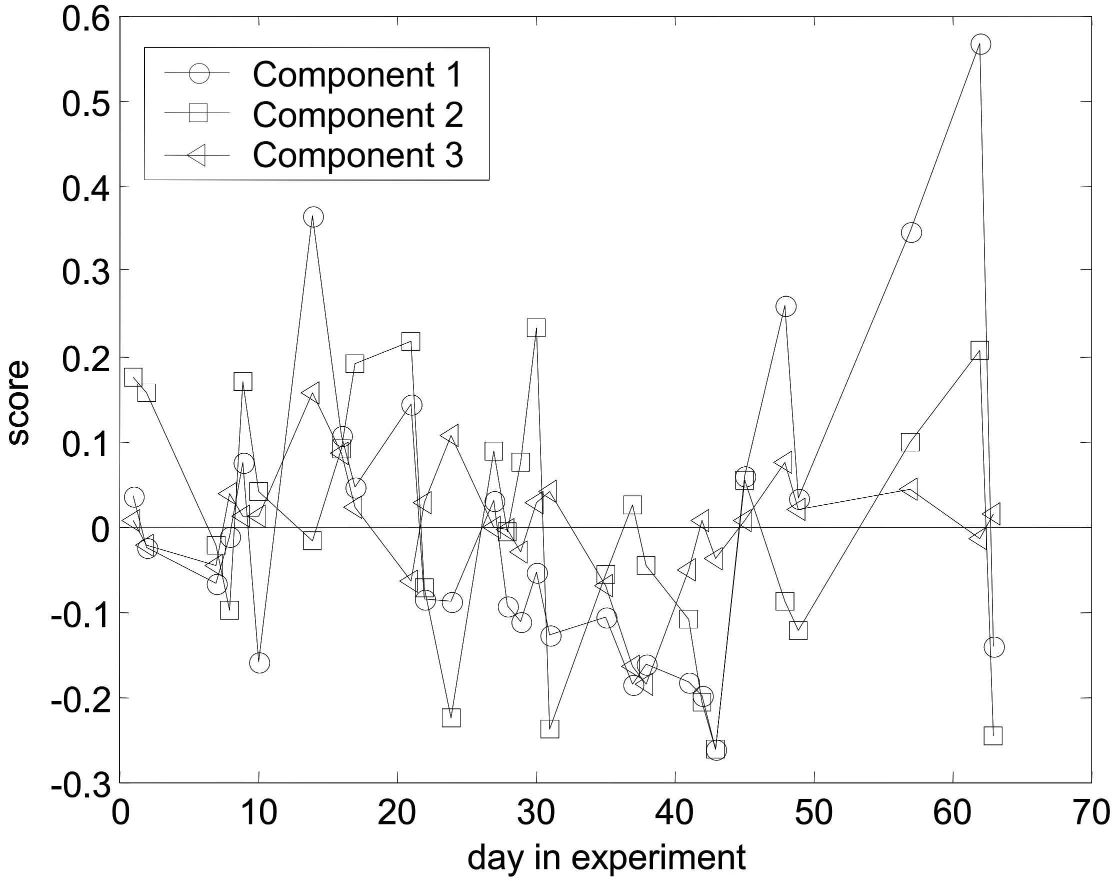

Figure 12 MSCA-P within-individual scores of monkey 6.

3.3.2 Results of the MSCA-P model

The results of the analysis of the monkey urine using MSCA-P presented here, should be seen as an illustration of the MSCA method.

The numbers of components for the between and the within-individual model are selected using a scree-graph. Three components are chosen for both the between and the within-individual models. The model explains 66 % of the within-individual variation and 77 % of the between-individual variation. The percentage of total variation in the data explained by the MSCA-P model is 68 %.

The within-individual scores of monkey 6 (a male monkey that is representative for the dataset) are given in Figure 12. These scores describe the dynamic variation of monkey 6 better than the corresponding PCA scores. The first component describes 45 % of the within-individual variation over all individuals. Components 2 and 3 describe 14 and 6 % respectively. The percentage of explained within-individual variation per component *for all individuals* is monotonously decreasing for increasing component number. This is not necessarily the case for the explained variation for each of these components *for each separate individual*. The percentage of explained variation of the first within-individual component for (female) monkey 10 is 34 %. For components 2 and 3 this is 13 and 16 % respectively. This means that the variation of monkey 10 on component 3 is larger than the variation on component 2. However, the variation over *all individuals* is larger in the direction of component 2 than in the direction of component 3. This means that the components of the within-individual model describe the most important metabolic biorhythms considering *all* individuals.

The within-individual variation explained per individual is between 49 and 77 %. The within-individual model of monkey 9 has an explained variation of 49 %, which is considerably less than the explained within-individual variation for the other monkeys. Therefore, there are differences between the metabolic biorhythms of monkey 9 and the other monkeys.

Figure 13 shows the between-individual scores of the male and female monkeys on the first and second components. The first between-individual component explains 33% of the between-individual variation and the second component explains 29%. Although this data analysis does not focus on gender differences, a separation between the scores of the male and female monkeys can be observed on the first component. This shows that the most important non-dynamic variation between the monkeys is related to gender differences.

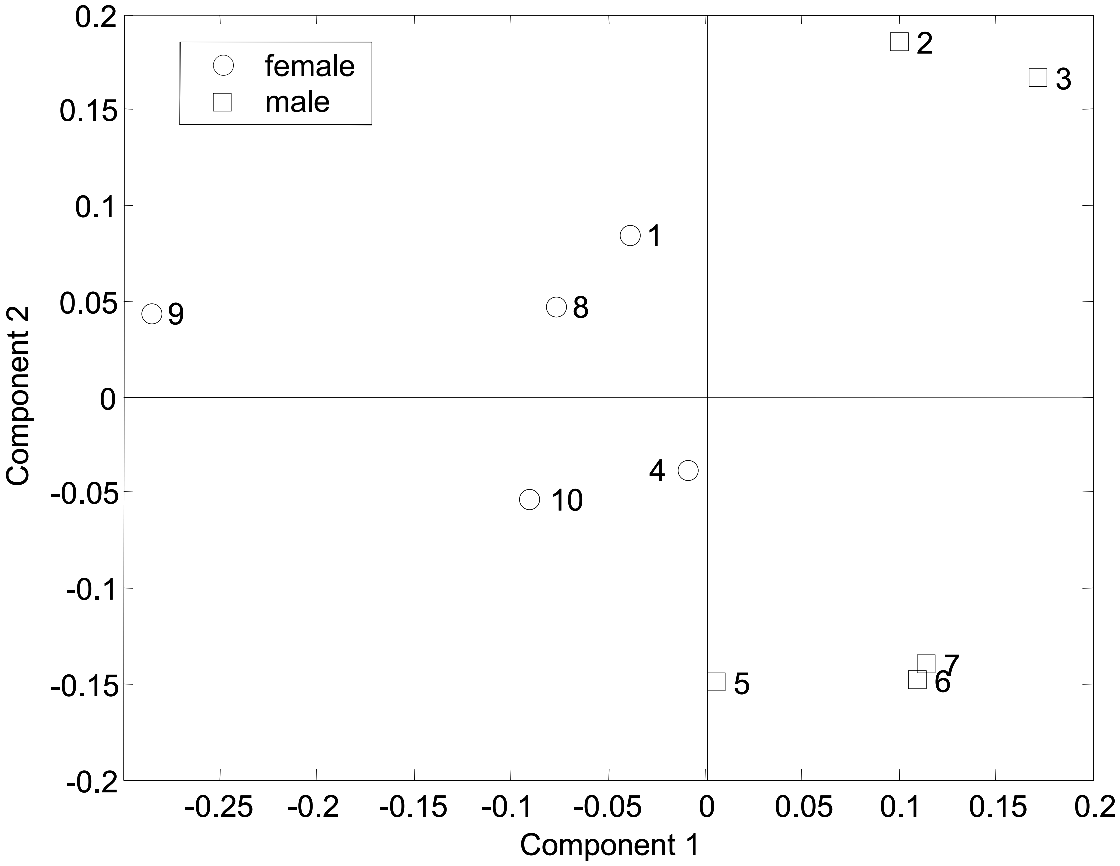

Figure 13 MSCA-P between-individual scores of the first and second component. A gender-based difference is visible on Component 1.

The loadings of the between and within-individual model can be compared to identify compounds that are important to describe the variation between and within the individuals. The loadings of the first component of the between-individual model are given in Figure 14. The loadings of the first component of the within-individual model are given in Figure 15. Comparison of both loadings shows that many compounds that are important to describe the variation on the within-individual level, are also important to describe the variation on the between-individual level (e.g. the loadings at 3.28, 3.05 and 1.93 ppm are large both in the within and between-individual model, these peaks can be attributed to trimethylamine N-oxide (TMAO), creatine and acetate respectively). There are also chemical shifts that have only a high loading for the between-individual model (e.g. 5.07 ppm, which cannot be assigned to a chemical compound and to a lesser extent 3.62 ppm, which can be attributed to fructose). This means that the concentration of these compounds varies between the monkeys, but is relatively constant in time within each monkey. It could be concluded that relatively large variation of these compounds in time for a certain monkey indicates abnormal behavior. This conclusion could not have been drawn from an interpretation of a PCA model of the monkey urine data, since in this model the between and within-individual variation are confounded.

Some compounds have a high loading for the within-individual model, but not for the between-individual model (e.g. the doublet at 1.32 and 1.34 ppm, that can be attributed to lactate). This means the concentration of lactate varies in time, but the average concentration of this compound is similar for different monkeys.

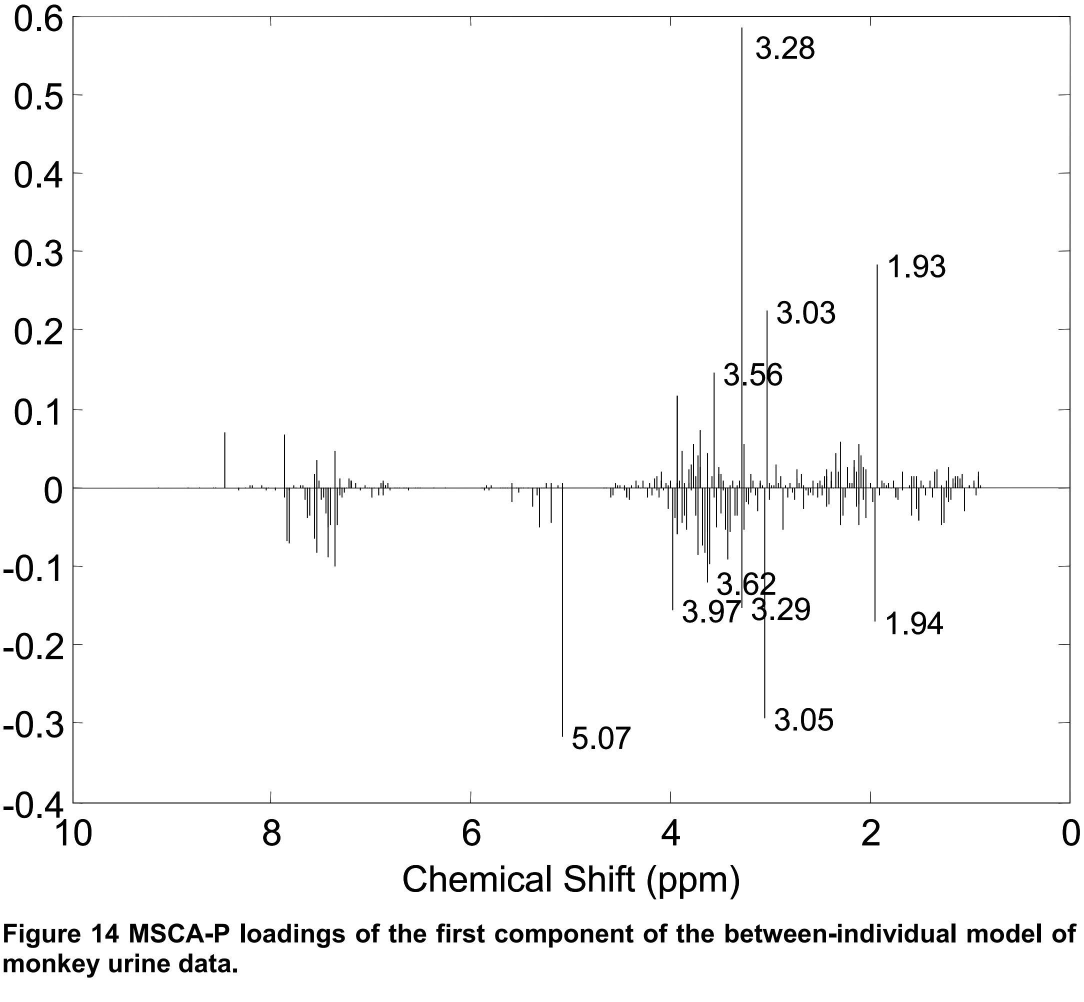

Figure 14 MSCA-P loadings of the first component of the between-individual model of the monkey urine data.

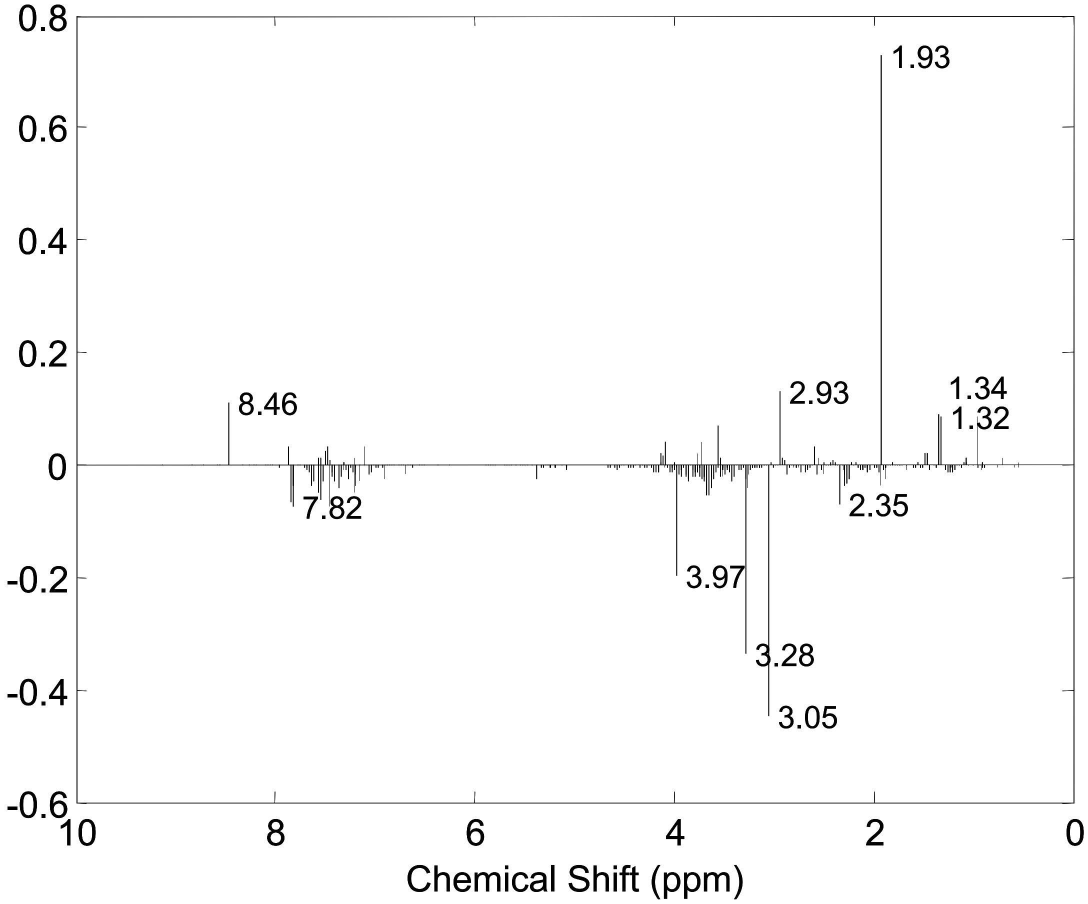

Figure 15 MSCA-P loadings of the first component of the within-individual model of the monkey urine data.

3.3.3 Comparison of the PCA and MSCA-P models

A PCA model is fitted to the monkey urine data using equation (2). The number of principal components for the model is determined by a scree-graph and 3 principal components are chosen. This PCA model describes 62 % of the variation in the data (the percentage of explained variation of PCA is lower than the percentage explained by MSCA-P, however the PCA model contains less components than the MSCA-P model).

To illustrate the distinction between the results of PCA and MSCA-P, the difference between the PCA scores and the MSCA-P within-individual scores of monkey 6 is calculated. This difference is given in Figure 16. This figure shows that the difference between the PCA and the within-individual MSCA-P scores of the first component is almost constant in time: the difference between both scores consists mainly of a (monkey-specific) offset. The differences between the scores of PCA and MSCA-P for the second and third components are clearly non-constant: for components 2 and 3 both methods describe different features in the data. The difference between the PCA and MSCA scores shows that the PCA model clearly does not only focus on the dynamic variation in the data: therefore it does not focus on describing the metabolic biorhythms.

Another way to compare the results obtained from the MSCA-P and PCA models is to focus on the non-dynamic variation in the data. This can be done by comparing the between-individual scores obtained from MSCA-P given in Figure 13 with the mean of the 29 PCA scores belonging to each monkey. The average PCA scores computed over all occasions for each individual for PC 1 and PC 2 are given in Figure 17. The average scores in Figure 17 show a separation between the male and the female monkeys on PC 2.

In the MSCA between-individual scores given in Figure 13, the gender-based differences could be straightforwardly interpreted to be responsible for the largest static variation between the monkeys. Due to the fact that in the PCA model different sources of variation are confounded, the gender-based differences are not identified as a factor underlying the largest variation in the data (because the PCA model also explains the dynamic variation in the data). This means that although the gender-based differences can also be observed from the PCA scores, the results of the MSCA model are much easier to interpret than the PCA model.

3.4 Conclusions

Time-resolved multisubject multivariate datasets (like those often acquired in metabolic fingerprinting) contain multiple types of variation. Methods that focus on all variation in the data simultaneously (like PCA), give a view on the data in which these different types of variation are confounded.

MSCA is a method that constructs a model that contains submodels in which different types of variation are described independently. A 2-level MSCA model contains a within-individual and a between-individual submodel.

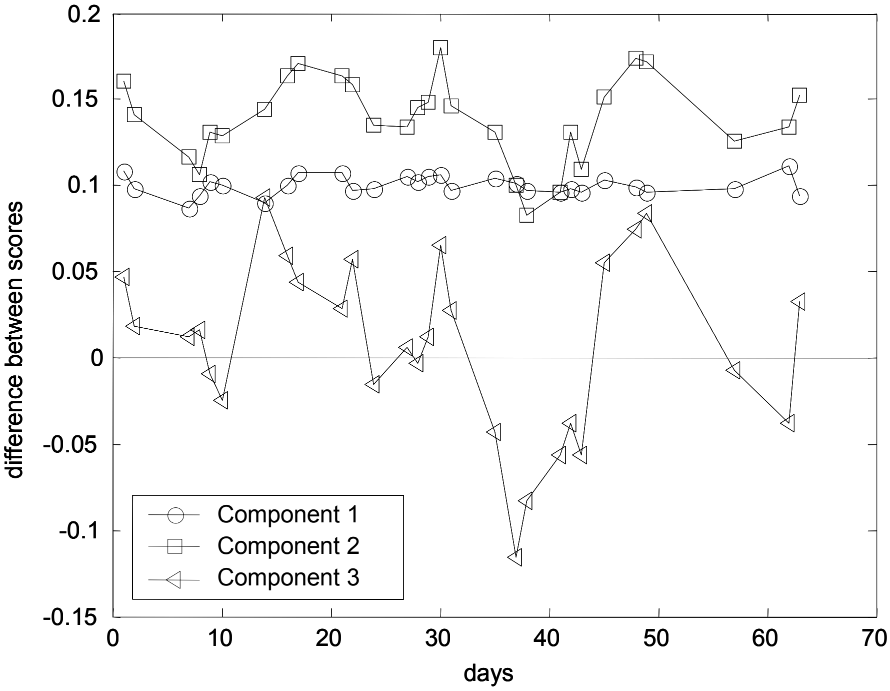

Figure 16 Difference between PCA and MSCA scores for monkey 6.

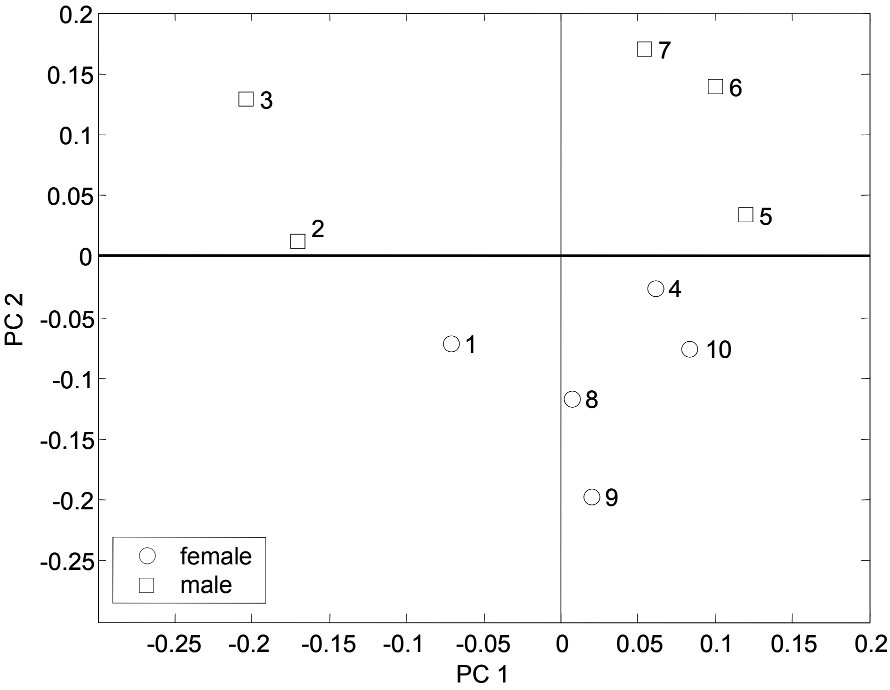

Figure 17 Average PCA scores for each monkey on PC 1 and PC 2. A gender-based difference is visible on PC 2.

The within-individual model gives a better representation of the dynamic variation in the data than PCA and the between-individual model gives a better representation and interpretation of the non-dynamic variation between the individuals than PCA. The price to pay for this improved interpretation is the fact that MSCA requires more components to be fitted for the same percentage of explained variation in the data as PCA.

The comparison of MSCA to PCA shows that MSCA leads to a better interpretation of all of the variation in a time-resolved metabolic fingerprinting dataset. However, scientific questions often refer to only one type of variation. When this type of variation is present in a multilevel dataset, MSCA can be also used to isolate this specific type of variation. Information about the other types of variation in the data is then not lost, since it is present in the other submodels of the MSCA model.

If the separation of within and between-individual variation is made in the analysis of the monkey urine data using a 2-level MSCA-P model, information can be obtained about the variation underlying the non-dynamic differences between the monkeys. The largest variation between the monkeys is identified to be gender-related. Furthermore, the dynamic variation in the urine composition of each monkey can be identified, thereby focusing on the normal metabolic biorhythms in the data. Comparison of the PCA and MSCA models of the same data shows that the different types of variation that are present in the data can be better interpreted using MSCA model than using PCA.

3.5 Acknowledgments

I would like to thank Robert-Jan Lamers for the chemical interpretation of the NMR-data and the obtained models.

3.6 Appendix: Notation

		Matrices and vectors
$\mathbf{X}_{\mathrm{raw}}$	$(L \times J)$	Raw data
$\mathbf{X}$	$(L \times J)$	Column centered data

$\hat{\mathbf{X}}$	$(L \times J)$	Estimated data of a subspace model of $\mathbf{X}$
$\mathbf{X}_i$	$(K \times J)$	Partition of $\mathbf{X}$ of individual i
$\mathbf{X}_{\mathbf{c},i}$	$(K \times J)$	Mean-centered $\mathbf{X}_i$
$\mathbf{X}_{\mathbf{c}}$	$(L \times J)$	Concatenation of all $\mathbf{X}_{\mathbf{c},i}$
$\mathbf{m}$	$(J \times 1)$	Vector containing column means of $\mathbf{X}_{\mathbf{raw}}$
$\mathbf{m}_i$	$(J \times 1)$	Vector containing column means of $\mathbf{X}_i$
$\mathbf{M}$	$(I \times J)$	Concatenation of all vectors $\mathbf{m}_i^{\mathbf{T}}$
$\mathbf{T}_{\mathbf{PCA}}$	$(L \times R)$	PCA scores
$\mathbf{t}_{\mathbf{b},i}$	$(R_{\mathbf{b}} \times 1)$	Between-Individual scores of individual i
$\mathbf{T}_{\mathbf{b}}$	$(I \times R_{\mathbf{b}})$	Between-Individual scores
$\mathbf{T}_{\mathbf{w},i}$	$(K \times R_{\mathbf{w}})$	Within-individual scores of individual i
$\mathbf{T}_{\mathbf{w}}$	$(L \times R_{\mathbf{w}})$	Within-individual scores of MSCA (concatenation of all $\mathbf{T}_{\mathbf{w},i}$)
$\mathbf{P}_{\mathbf{PCA}}$	$(J \times R)$	PCA loadings
$\mathbf{P}_{\mathbf{b}}$	$(J \times R_{\mathbf{b}})$	Between-Individual loadings
$\mathbf{P}_{\mathbf{w},i}$	$(J \times R_{\mathbf{w}})$	Within-individual loadings of individual i
$\mathbf{P}_{\mathbf{w}}$	$(J \times R_{\mathbf{w}})$	Within-individual loadings of MSCA
$\mathbf{E}$	$(L \times J)$	Model residuals
$\mathbf{E}_{\mathbf{PCA}}$	$(L \times J)$	PCA model residuals
$\mathbf{E}_{\mathbf{MCA}.i}$	$(K_i \times J)$	MCA model residuals of individual i
$\mathbf{E}_{\mathbf{MSCA}.i}$	$(K_i \times J)$	MSCA model residuals of individual i
$\mathbf{E}_{\mathbf{b}}$	$(I \times J)$	Between-individual model residuals
$\mathbf{E}_{\mathbf{w}}$	$(L \times J)$	Within-individual model residuals

$\mathbf{1}_L$, $\mathbf{1}_{K_i}$, $\mathbf{1}_I$ Size L, K_i, I column vector containing ones

$\mathbf{0}_{R_\mathbf{b}}$, $\mathbf{0}_{R_{\mathbf{w},i}}$ Size $R_{\mathbf{w},i}$, $R_\mathbf{b}$ column vector containing zeros

Scalars

L Number of spectra in the dataset $= \sum_{i=1}^{I} K_i$

$1 \ldots i \ldots I$ Index for the individuals

J Number of variables (chemical shifts)

K_i Number of measurement time-points (occasions) of individual i

K Number of measurement time-points

R Number of principal components of the PCA model

$R_{\mathbf{w},i}$ Number of principal components of the within-individual MCA model for individual i

$R_\mathbf{b}$ Number of principal components of the between-individual MSCA model

$R_\mathbf{w}$ Number of principal components of the within-individual MSCA model

$f_\mathbf{MSCA}$ Least squares criterion of MSCA

3.7 References

[1] J. van der Greef, A. C. Tas, J. Bouwman, M. C. Ten Noever de Brauw, W. H. P. Schreurs, Anal.Chim.Acta, 150 (1983) 45

[2] K. P. Gartland, S. M. Sanins, J. K. Nicholson, B. C. Sweatman, C. R. Beddell, J. C. Lindon, NMR Biomed., 3 (1990) 166

[3] Glass, L. and Mackey, M. C., From Clocks to Chaos: "The Rythms of Life", Princeton University Press, Princeton, 1988

[4] A. C. Tas, H. van den Berg, J. Odink, H. Korthals, J. T. N. M. Thissen, J. van der Greef, J.Pharmaceut.Biomed., 7 (1989) 1239

[5] E. Holmes, P. J. D. Foxall, J. K. Nicholson, G. H. Neild, S. M. Brown, C. R. Beddell, B. C. Sweatman, E. Rahr, J. C. Lindon, M. Spraul, P. Neidig, Anal.Biochem., 220 (1994) 284

[6] C. Zuppi, I. Messana, F. Forni, C. Rossi, L. Pennacchietti, F. Ferrari, B. Giardina, Clin.Chim.Acta, 265 (1997) 85

[7] J. R. Bales, D. P. Higham, I. Howe, J. K. Nicholson, P. J. Sadler, Clin.Chem., 30 (1984) 426

[8] J. G. Bundy, E. M. Lenz, D. Osborn, J. M. Weeks, J. C. Lindon, J. K. Nicholson, Xenobiotica, 32 (2002) 479

[9] M. E. Dumas, C. Canlet, F. Andre, J. Vercauteren, A. Paris, Anal.Chem., 74 (2002) 2261

[10] S. Garrod, E. Humpher, S. C. Connor, J. C. Connelly, M. Spraul, J. K. Nicholson, E. Holmes, Magn.Res.Med., 45 (2001) 781

[11] J. K. Nicholson, J. A. Timbrell, P. J. Sadler, Mol Pharmacol, 27 (1985) 644

[12] E. Holmes, H. Antti, The Analyst, 12 (2002) 1549

[13] J. J. Jansen, H. C. J. Hoefsloot, H. F. M. Boelens, J. van der Greef, A. K. Smilde, (In Press), Bioinformatics

[14] H. Antti, M. E. Bollard, T. M. Ebbels, H. Keun, J. C. Lindon, J. K. Nicholson, E. Holmes, J.Chemometr., 16 (2002) 461

[15] P. Mendes, Briefings in Bioinformatics, 3 (2002) 134

[16] Snijders, T. and Bosker, R., Multilevel Analysis, Sage Publications Ltd., London, 1999

[17] H. A. L. Kiers, J. M. F. Ten Berge, Br.J.Math.Stat.Psychol., 47 (1994) 109

[18] M. E. Timmerman, H. A. L. Kiers, Psychometrika, 86 (2003) 105

[19] M. E. Timmerman, Br. J Math. Stat. Psy., in press

[20] J. T. W. E. Vogels, A. C. Tas, J. Venekamp, J. van der Greef, J.Chemometr., 10 (1996) 425

[21] Jolliffe, I. T., Principal Component Analysis, Springer-Verlag, New York, 2002

[22] R. Bro, A. K. Smilde, J.Chemometr., 17 (2003) 16

[23] J. B. Kruskal, Manuscript, Bell laboratories, Murray Hill, NJ (1977)

[24] Ten Berge, J. M. F., Least squares optimization in multivariate analysis, DSWO press, Leiden, 1993

[25] Jackson, J. E., A User's Guide to Principal Components, (Wiley Series in probability and mathematical statistics), Wiley-Interscience, New York, 1991

[26] Searle, S. R., Casella, G., and McCullough, C. E., Variance Components, (Wiley Series in Probability and Statistics), John Wiley & Sons, Inc., 1992

[27] R. B. Cattell, Multivar.Behav.Res., 1 (1966) 245

[28] H. A. L. Kiers, J. M. F. Ten Berge, R. Bro, J.Chemometr., 13 (1999) 275

4 ANOVA-Simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data

4.1 Introduction

Recent developments in genomics and human systems biology have shown the importance of metabolomics (Clish, Davidov *et al.*, 2004; Lindon, Holmes *et al.*, 2004; van der Greef, Davidov *et al.*, 2003). This is understandable, since metabolomics is a crucial element in bridging the difference between the genotype and phenotype of an organism (Fiehn, 2002). Considerable effort has gone into the development of instrumental methods for metabolite profiling, since it emerged in the sixties and seventies in clinical chemistry (Gates and Sweeley, 1978; Jellum, 2001), specially focusing on inborn errors of metabolism. The combination with mass spectrometry (GC-MS) and in particular chemometrics created the basis of the technology of today in metabolomics for mammalian systems (Gaspari, Vogels *et al.*, 2001; van der Greef, Tas *et al.*, 1983). Nuclear Magnetic Resonance (NMR) spectroscopy has become an important component in the field for screening of biofluids (Bales, Higham *et al.*, 1984) and an important step was the combination with chemometrics (Gartland, Sanins *et al.*, 1990). Moreover, methods for handling metabolomics data are receiving increased attention, both regarding preprocessing of metabolomics data (Keun, Ebbels *et al.*, 2003; Vogels, Tas *et al.*, 1996) as well as in the analysis of data itself (Antti, Bollard *et al.*, 2002; Jansen, Hoefsloot *et al.*, 2004; Keun, Ebbels *et al.*, 2004).

Metabolomics data sets are becoming more and more complex. It is not uncommon to measure a multiple of metabolites in body fluids of several animals, at different points in time with an underlying experimental design, e.g. different dose groups (Antti, Ebbels *et al.*, 2004; Keun, Ebbels *et al.*, 2004; Lamers, DeGroot *et al.*, 2003). This calls for data analysis methods specifically suited for time-resolved (or ‘longitudinal’), multigroup, multisubject (containing data of multiple animals) and multivariate data.

For the case of measuring a single variable (e.g. a metabolite) as a function of design factors, analysis of variance (ANOVA) is a well established technique to analyze the data (Searle, 1971). When measuring many metabolites simultaneously, generalizations of ANOVA are necessary. In the statistics literature, the classical generalization of ANOVA to multiple variables is multivariate-ANOVA (MANOVA) (Mardia, Kent *et al.*, 1979). For the large number of measured variables in a metabolomics experiment, however, MANOVA breaks down due to problems of singularity of covariance matrices and assumptions that are not fulfilled (Ståhle and Wold, 1990).

In the data analysis literature, also mixtures of multivariate analysis and ANOVA have been reported. One approach first performs a principal component analysis of the whole data set and then uses ANOVA on the component score values to test effects (Bratchell, 1989). This approach has been criticized, since the separate ANOVA's on the score values are not independent (Jackson, 1991). Moreover, the initial principal component analysis does not necessarily distinguish between the groups in the data. Another approach is suggested by using PLS (partial least squares; a popular regression technique for collinear data) to solve the problem (Ståhle and Wold, 1990). However, different suggestions have been made to implement this method depending on whether the coded design variables are regressors (Martens and Martens, 2001) or regressands (Ståhle and Wold, 1990). Moreover, the exact properties of these methods are unknown, since the criterion that is maximized or minimized is not clear. Redundancy analysis using a coded design matrix seems to be a better alternative than the PLS based approach (Van den Brink and Ter Braak, 1999).

In this paper, a new method is presented that can deal with a temporal and/or design structure of complex multivariate data sets such as those emerging from metabolomics experiments. However, the problems to which ASCA can be applied are certainly not limited to metabolomics. Complex multivariate datasets are abundant in the other postgenomic technologies (e.g. transcriptomics, proteomics) and also in many more fields of biological and non-biological research.

ASCA builds upon and generalizes some earlier proposed methods. Two early papers in pomology and botanics realized the importance of distinguishing *between* and *within* factor treatments (Jeffers, 1962; Pearce and Holland, 1960). In the metabolomics literature, the SMART method (Keun, Ebbels *et al.*, 2004) also makes this distinction, but is less general than the method proposed in the current paper. The method proposed in this paper builds on the Multilevel Component Analysis method that was developed in psychometrics (Timmerman, In Press) and metabolomics (Jansen, Hoefsloot *et al.*, 2004) and generalizes it for a situation with any design structure underlying the metabolomics data.

The ANOVA-SCA (ASCA) method will be explained and illustrated with an example from a metabolomics intervention study, where guinea pigs from a strain developing osteoarthritis (OA) were treated with several dosage levels of vitamin C and their urine was analysed using NMR spectroscopy at several points in time. This study is therefore a typical example of a designed metabolomics experiment. Osteoarthritis is a multi-factorial chronic joint disease that is characterized by the progressive destruction of articular cartilage, resulting in impaired movement, pain and ultimately disability (Creamer and Hochberg, 1997). The Hartley outbred strain guinea pig develops spontaneous progressive knee osteoarthritis starting when they are about 10 months of age, with features similar to the human disease (Bendele, 2001; Huebner, Otterness *et al.*, 2001). Ascorbic acid has been associated with the slowing of osteoarthritis progression in guinea pig and human (McAlindon, Jacques *et al.*, 1996). However, recent studies indicate that Vitamin C increases the severity of development of OA in the guinea pig (Kraus, Huebner *et al.*, 2004). The details of the biological questions regarding this study are published elsewhere (Lamers, DeGroot *et al.*, 2003).

4.2 System and Methods

4.2.1 Urine samples and data acquisition

This dataset contains information about Male Hartley guinea pigs, which develop osteoarthritis during aging. Beginning at 4 months of age, the guinea pigs are divided randomly into three dose groups to which varying doses of vitamin C are provided: low dose (2.5–3 mg/day), medium dose (30 mg/day), and high dose (150 mg/day). The doses were chosen such, that the low dose exceeds the minimum amount to prevent scurvy and the medium dose corresponds to the normal intake of Vitamin C. The high dose of Vitamin C corresponds to the amount that was shown in previous studies to slow the development of surgically induced OA.

Each dose group consists of 6 animals. Urine samples are collected at 4, 7, 10 and 12 months, where the samples collected after 4 months are pre-dose. Each urine collection was performed for 24 hours, to remove the influence of diurnal variation of the metabolite composition of the urine. The total dataset consists of 72 samples. These samples were analyzed with NMR spectroscopy and the dataset was prepared as peak listings (NMR spectra) using the standard Varian software (Varian inc., Palo Alto, CA). The dataset contains spectra with peaks listed at 253 chemical shifts, expressed in parts per million (ppm), that are equal for all spectra. A typical spectrum in this dataset is given in Figure 18.

The dataset is a subset of a larger dataset. The acquisition of this larger dataset has been described elsewhere (Lamers, DeGroot *et al.*, 2003).

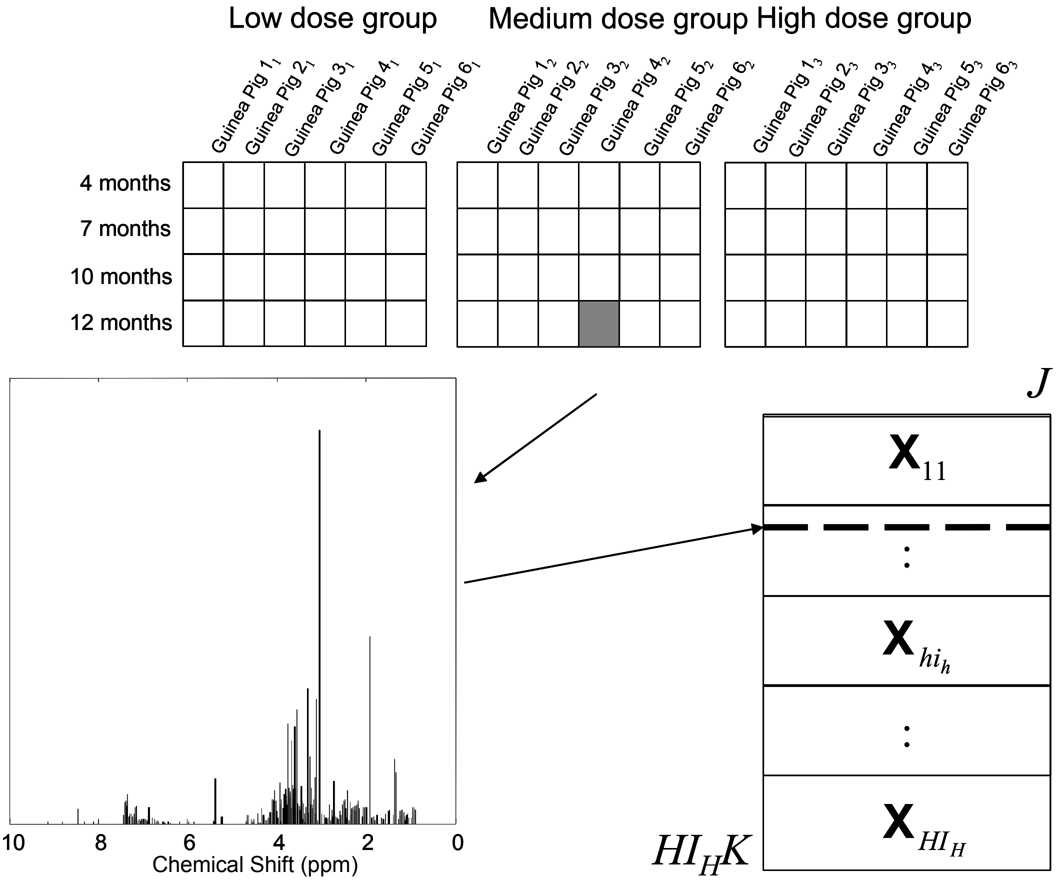

Figure 18 Structure of the dataset: the relationship between the measurements is given in the top of the figure: the guinea pigs are nested within the dose groups, all other relationships between the factors in the experiment are crossed. Each square in the top of the figure represents a NMR spectrum like the one given in the bottom left corner. These obtained spectra are arranged into a matrix containing HIH submatrices, as indicated in the bottom right corner.

4.2.2 Data analysis

4.2.2.1 Structure of the data set

The structure of the data set is shown in Figure 18. The following indices will be used: $j=1,...,J$ for the chemical shifts; $k=1,...,K$ for the time-points at which measurements are taken; $h=1,..,H$ for the dosage groups (*h*=1: low; *h*=2: medium and *h*=3: high dosage); $i_h=1,...,I_h$ for the guinea pigs nested within the dosage groups: the guinea pigs in dosages group 'low' are different from those in the other dosage groups (the subindex h on i is used to stress this fact). This will be important to realize for the remainder of the paper.

4.2.2.2 Analysis of variance

An NMR signal at one particular chemical shift *j*, for one time point k, and for one guinea pig i_h (in dosage group h) will be denoted by the scalar x_{hki_hj}. Collecting such values x_{hki_hj} in matrices $\mathbf{X}_{hi_h}$ of size (K x J) will be convenient. The construction of these matrices is shown in Figure 18. Considering only an NMR signal at one chemical shift (and therefore dropping the index j for convenience) a reasonable ANOVA model would be

$$x_{hki_h} = \mu + \alpha_k + (\alpha\beta)_{hk} + (\alpha\beta\gamma)_{hki_h} \tag{1}$$

where μ represents an overall offset; α_k represents the effect of the factor 'time' that is common for all guinea pigs; $(\alpha\beta)_{hk}$ represents the interaction of 'time' and 'dose'; $(\alpha\beta\gamma)_{hki_h}$ represents the guinea pig specific contribution. Of these effects, $(\alpha\beta)_{hk}$ is most important for the biological interpretation: it represents the effect of the dosage measured as deviations from the common time effect α_k . The contributions $(\alpha\beta\gamma)_{hki_h}$ represent the variations on the lowest (individual animal-specific) level, and can be used for significance testing. Classical ANOVA techniques can now be used to estimate the factor effects and test significance.

Equation (1) shows a division of variation on several factors. This is the basic idea of ANOVA: variation is separated and assigned to factors. The factor effects can be estimated and tested. ANOVA is capable of doing this by splitting the variations in orthogonal and independent parts (Searle, 1971). This division of the variation into orthogonal contributions is also the goal of ANOVA-SCA (see below).

4.2.2.3 Simultaneous Component Analysis

When analyzing the simultaneous underlying variation in several related data sets, simultaneous component analysis is a useful tool. This method was developed in psychometrics (Ten Berge, Kiers *et al.*, 1992), but extensions also found their use now in metabolomics (Jansen, Hoefsloot *et al.*, 2004; Timmerman, In Press).

Suppose data matrices $\mathbf{X}_i$ (K_i x J) are available where measurements on J identical chemical shifts are available at K_i time-points on I animals (the subdivision of the individuals i into different dose groups h is omitted from the explanation of SCA for simplicity, therefore the used indices are simpler in this section compared to the other sections of this paper). Note that the number of measurement time-points for individual i, denoted by K_i, can differ between animals in SCA. Then a reasonable model for simultaneously analyzing these data matrices is

$$\mathbf{X}_i = \mathbf{T}_i\mathbf{P}^{\mathrm{T}} + \mathbf{E}_i \qquad (2)$$

where **P** of dimensions (J x R) represents the common basis with R directions (components) and $\mathbf{T}_i$ of dimensions (K_i x R) contains the scores of the measurement time-points of the ith animal. Since the variation across the animals at the various time-points is expressed on the common basis **P**, the scores contained in $\mathbf{T}_i$ can be compared between animals to explore the data. There exist different versions of simultaneous component analysis depending on the type of constraint put on the covariance of $\mathbf{T}_i$, but such constraints are not discussed in this paper.

The model parameters in equation (2) can be found by solving

$$\min_{\mathbf{T}_i\text{'}s,\mathbf{P}} \sum_{i=1}^{I} \left\| \mathbf{X}_i - \mathbf{T}_i\mathbf{P}^{\mathrm{T}} \right\|^2 \qquad (3)$$

which is a standard least squares problem that can be solved by performing a PCA on the matrix in which all matrices $\mathbf{X}_i$ are concatenated as $\begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_I \end{bmatrix}$.

4.2.2.4 ANOVA-Simultaneous component analysis (ASCA)

The ANOVA-SCA analog for model (1) is

$$\mathbf{X}_{hi_h} = \mathbf{1m}^{\mathbf{T}} + \mathbf{T_K P_1^T} + \mathbf{T}_{\mathbf{K}h}\mathbf{P_2^T} + \mathbf{T}_{\mathbf{K}hi_h}\mathbf{P_3^T} + \mathbf{E}_{hi_h} \tag{4}$$

where **1** is a $(K \times 1)$ vector of ones; **m** is a $(J \times 1)$ vector of the overall means of the NMR responses (where the mean is taken over all factors and guinea pigs per chemical shift); $\mathbf{T_K}$ of dimensions $(K \times R_1)$ is the matrix containing the contributions of the factor 'time' expressed on the basis $\mathbf{P_1}$ of dimensions $(J \times R_1)$; R_1 is the number of components chosen for the basis $\mathbf{P_1}$; $\mathbf{T}_{\mathbf{K}h}$ of dimensions $(K \times R_2)$ is the matrix containing the dose-specific 'time' contributions ($h=1,...,H$) expressed on the basis $\mathbf{P_2}$ of dimensions $(J \times R_2)$; R_2 is the number of components chosen for the basis $\mathbf{P_2}$; $\mathbf{T}_{\mathbf{K}hi_h}$ of dimensions $(K \times R_3)$ is the matrix of guinea pig specific dose-time contributions ($i_h = 1,...,I_h \forall h$) expressed on the basis $\mathbf{P_3}$ of dimensions (JxR_3); R_3 is the number of components chosen for the basis $\mathbf{P_3}$ and $\mathbf{E}_{hi_h}$ is the matrix of residuals. Note that in equation (4), $\mathbf{T_K}$ is equal for all animals, $\mathbf{T}_{\mathbf{K}h}$ is equal for all animals belonging to the same dose group and $\mathbf{T}_{\mathbf{K}hi_h}$ is different for all animals. Because in $\mathbf{T_K}$ and $\mathbf{T}_{\mathbf{K}h}$ each time-point K_{hi_h} is compared between different individuals, the measurement time-points are required to be equal for all animals, such that $K_{hi_h} = K$.

In words, equation (4) means that the matrix **X** is separated into contributions from the overall mean ($\mathbf{1m^T}$), one SCA model ($\mathbf{T_K P_1^T}$) describing the overall effect of the factor time, a SCA model ($\mathbf{T}_{\mathbf{K}h}\mathbf{P_2^T}$) describing the interaction of dose with time, and a SCA model ($\mathbf{T}_{\mathbf{K}hi_h}\mathbf{P_3^T}$), describing the interaction of dose, time and guinea pig, which is the contribution to the variation of each individual guinea pig. This is a direct multivariate generalization of equation (1). Note that since the number of components in each part of this model is low (which is the basic idea:

dimension reduction) there is a residual matrix $\mathbf{E}_{hi_h}$ in equation (4) that contains the information that is not described by any of the ASCA submodels, whereas such a residual term is not present in (1). To illustrate the basic idea behind ASCA, a toy example is given in Appendix 1 at the end of this chapter.

4.2.2.5 Properties of ASCA

By imposing the proper constraints, the different parts of the model in equation (4) are orthogonal to each other. These constraints are

$$
\begin{aligned}
&a)\,\mathbf{1}^{\mathrm{T}}\mathbf{T}_{\mathbf{K}} = \mathbf{0}^{\mathrm{T}} \\
&b)\sum_{h=1}^{H}\mathbf{T}_{\mathbf{K}h} = \mathbf{0} \\
&c)\sum_{i_h=1}^{I_h}\mathbf{T}_{\mathbf{K}hi_h} = \mathbf{0}\ \forall\ h = 1,\ldots,H
\end{aligned}
\qquad (5)
$$

where **1** is a vector of ones of the proper order and **0** is a vector or a matrix of zeros of the proper order. In words, a) ensures that $\mathbf{T}_{\mathbf{K}}$ is orthogonal to $\mathbf{1m}^{\mathrm{T}}$, and likewise b) and c) ensure orthogonality of the other parts (for a detailed description of the mathematics behind ASCA, see Appendix 2 at the end of this chapter). This also means that the total variation of the data set can be separated in parts corresponding to the different factors.

4.2.2.6 ASCA-Algorithms

The parameters of the ASCA model can be calculated by solving the following least squares problem

$$
\min_{\mathbf{m},\mathbf{T's},\mathbf{P's}} \sum_{h=1}^{H}\sum_{i_h=1}^{I_h}\left\| \mathbf{X}_{i_h} - \mathbf{1m}^{\mathrm{T}} - \mathbf{T}_{\mathbf{K}}\mathbf{P}_{\mathbf{1}}^{\mathrm{T}} - \mathbf{T}_{\mathbf{K}h}\mathbf{P}_{\mathbf{2}}^{\mathrm{T}} - \mathbf{T}_{\mathbf{K}hi_h}\mathbf{P}_{\mathbf{3}}^{\mathrm{T}} \right\|^2 \qquad (6)
$$

under the constraints given in equation (5). Although this looks like a complicated problem, these constraints actually make the problem relatively simple. Due to these contraints the parameter sets corresponding to the different factor combinations ($\mathbf{m}$; $\mathbf{T}_{\mathbf{K}}$, $\mathbf{P}_1$; $\mathbf{T}_{\mathbf{K}h}$, $\mathbf{P}_2$; $\mathbf{T}_{\mathbf{K}hi_h}$, $\mathbf{P}_3$) can be estimated independently. Note that there is rotational freedom in the model parts containing $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$. Hence, the matrices $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$ can be chosen to be orthogonal. For the case of a balanced design (equal number of guinea pigs for each factor combination) it comes down to proper centering and performing PCA's on rearranged data. An algorithm is provided, but standard algorithms for PCA can also be used after the proper centering and rearrangement of the data. For the unbalanced case, a slightly more elaborate algorithm should be used, which is a straightforward generalization of the balanced one. For an explanation of the algorithm, see the Appendix 3 given in the additional material.

4.3 Results

4.3.1 Split-up of variation

Table 1 Contributions to the total variation

Level	% of variation in the data
K	24
Kh	10
Khi_h	66

An impression of the amount of variation related to the design factors can be obtained by separating this variation into contributions from the different factors. Table 1 shows this separation and it is clear that the dominant part of the variation is at the lowest level (guinea pig-specific contributions). This shows the biological variation between the animals used in the study. Note that Table 1 reports sums of squared deviations from the overall mean and not variances.

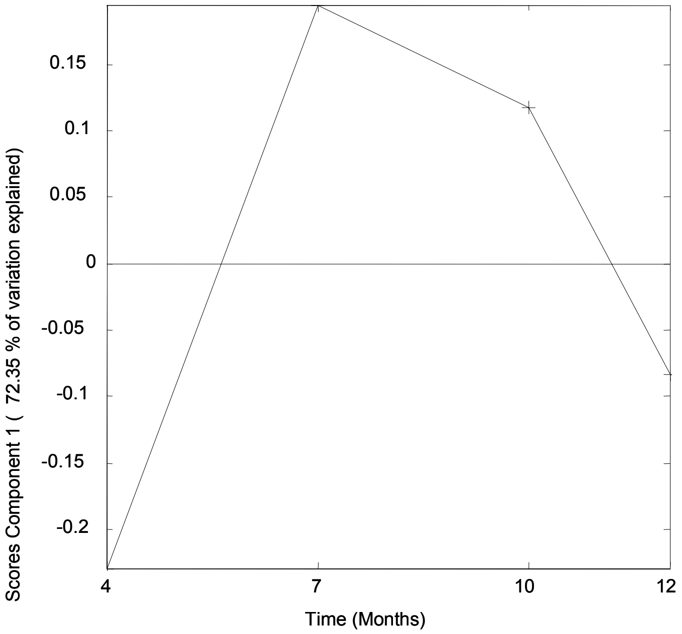

Figure 19 factor 'Time' scores on the first component: clearly an initially increasing and subsequently decreasing time profile is visible.

4.3.2 Factor 'time'

The scores of the first component of the factor 'time' are given in Figure 19 (the $\mathbf{T}_K$ values of Equation (4)). This first component explains 72 % of the variation on the factor *'time'*. The maximum number of components that can be fitted for the factor *'time'* submodel (and therefore the rank of the factor time variation) is 3, since the dataset contains 4 measurement time-points. Hence, only one component is used to illustrate the variation on this level.

The scores in Figure 19 indicate that all guinea pigs in the data show an initial increasing and a subsequent decreasing behavior. This trajectory is consistent with the biology of growth for the Hartley guinea pig strain (Huebner, Otterness *et al.*, 2001). From 4 to 7 months the metabolism of the guinea pigs changes, because during this time they are in the growing phase. Between 7 and 10 months, the guinea pigs are full-grown, which is shown by the leveling off in the time profile from 7 and 10 months. From 10 months on the guinea pigs develop osteoarthritis. The decrease in score of the 12 months samples is supposed to reflect this effect (Lamers, DeGroot *et al.*, 2003).

The loadings $\mathbf{P}_1$ belonging to the first component are given in Figure 20. These loadings show the chemical shifts and therefore the compounds that are corresponding to the behavior observed in the scores and can be used for biological interpretation. However, urine is a biological fluid that is used by the body for the excretion of waste products and therefore its contents are difficult to trace back to biology. Nevertheless, metabolites like creatinine (δ 3.04 and 4.05 ppm), creatine (δ 3.04 and 3.95 ppm), glucose (δ 3.27, 3.53, 3.60, 3.78, 3.82 and 3.94 ppm) and alpha-hydroxybutyrate (δ 1.36 ppm), lactate (δ 1.32 and 1.34 ppm), glycine (δ 3.56 ppm) and acetate (δ 1.92 ppm) that change in time may point at altered energy metabolism due to growth and disease development. These observations are consistent with results that were described previously (Lamers, DeGroot *et al.*, 2003).

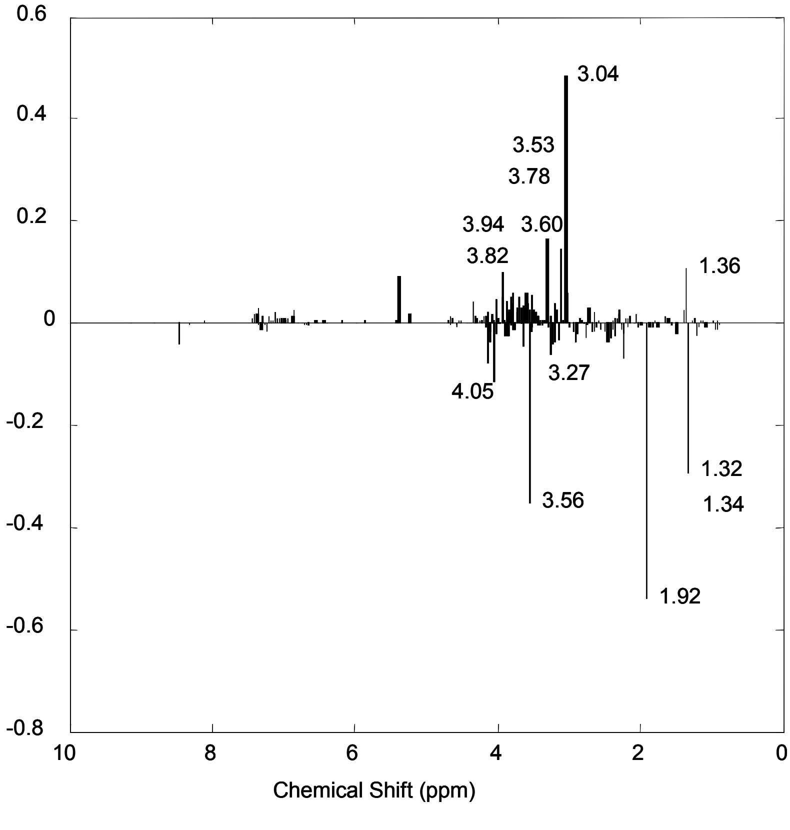

Figure 20 factor 'Time' loadings on the first component. The chemical shifts that are mentioned in the text are indicated in the figure

4.3.3 Interaction time x dose

The rank required for this submodel is determined using a scree-test (Cattell, 1966). From this test, the rank of this submodel is determined to be two. A model containing two components explains 65 % of the variation corresponding to the factor '*interaction time x dose'.* The first component explains 50 % of this variation and the second component explains 15 %.

The scores of this submodel should be seen be interpreted as the deviation of each dose group from the 'time' factor (($\mathbf{T_K P_1^T}$ or α_k in the ANOVA model). The 'interaction time x dose' scores for the first component are given in Figure 21. On this component there is no trend related to Vitamin C dose visible: none of the 4 measurement time-points show an increasing or decreasing score value for differing Vitamin C doses. Also the scores of the second component do not show such a quantitative trend.

The model results show that the potential of vitamin C in affecting the development of osteoarthritis is questionable. According to our results vitamin C has no effect on disease development: neither the association of vitamin C with the slowing of osteoarthritis progression in guinea pig and humans (McAlindon, Jacques et al., 1996) nor the observation that vitamin C could increase the severity of development of OA in the guinea pig (Kraus, Huebner et al., 2004) can be corroborated with the results of ASCA on the described dataset. However, the results of the ASCA model of this dataset were in agreement with additional clinical measurements that were performed after the experiment on the guinea pigs used in the study: the severity of OA was determined using 'histology scores' on the Mankin grading system (Mankin HJ, Dorfman H, et al., 1971). These scores did not differ between dose groups that indicates an equal development of OA.

4.3.4 Individual guinea pig contributions

The rank required for the *'Individual guinea pig contributions'* submodel can also be determined by a scree-test. From this test, two components are defined for this submodel. This submodel describes 57 % of the variation corresponding to the *'Individual guinea pig contributions'*. The first component explains 45 % and the second component explains 12 %. The scores of this submodel must be seen as a deviation of each individual from the dose-time interaction.

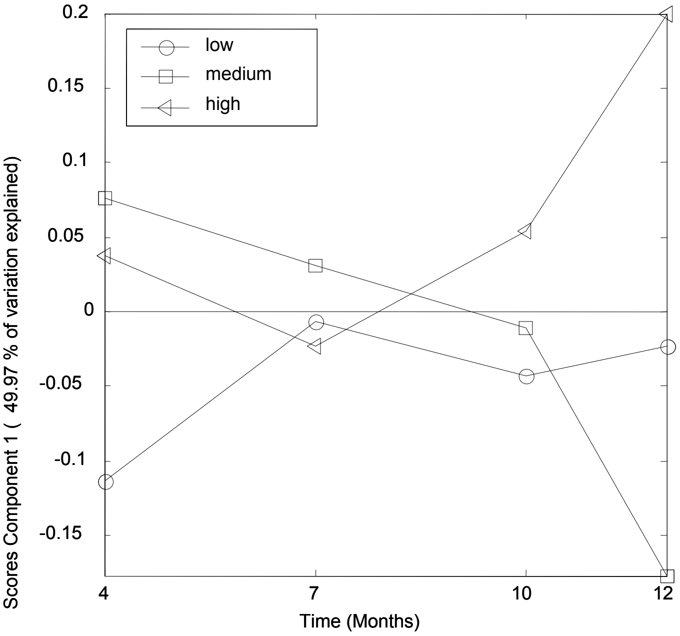

Figure 21 Interaction 'Dose x Time' scores on the first component. No quantitative effect is visible in the scores and therefore this model shows that Vitamin C has no effect on the development of OA.

The scores for the first component of the *'Individual guinea pig contributions'* submodel are given in Figure 5 for the low, medium and high dose groups respectively. From this figure it is clear that the deviation of the individual profile from the group average profile is largest at 4 and 12 months: at the start and the end of the experiment. The NMR signals that correspond to this behavior are, amongst others, lactate (δ 1.32 and 1.34 ppm), acetate (δ 1.92 ppm) and glycine (δ 3.56 ppm) that increase and creatinine (δ 3.04 and 4.05 ppm) that decrease. The larger interindividual differences at 4 and 12 months may be explained by the fact that at these timepoints growth and disease development, respectively, occurs.

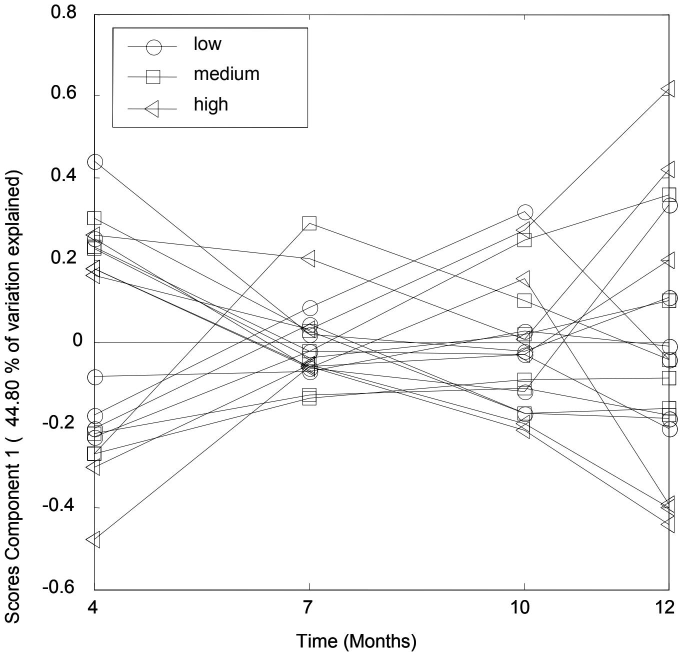

Figure 22 'Individual guinea pig contribution' scores on the first component. These scores indicate the deviation of each individual from the dose-group specific variation of the metabolism.

4.4 Conclusions

In metabolomics, an increasing amount of data sets becomes available with an underlying design in factors. Currently, no methods are available to analyze such data. The method proposed in this paper called ANOVA-SCA or ASCA for short, fills this gap. It works by separating the variation in the total data set by parts that can be assigned to contributions of the different factors and interactions thereof.

The ASCA method is illustrated by a real example of an intervention study examining the effect of Vitamin C on the development of osteoarthritis in guinea pigs. This shows how the method works and that interpretation of the resulting components works in the same way as in ordinary principal component analysis. For the case of a balanced design, the algorithm is simple and comes down to performing principal component analyses on properly centered and rearranged data. For the case of unbalanced data, a more elaborate algorithm is necessary, but available.

In follow-up research, questions regarding validation using resampling methods and significance testing of effects will be treated. This will allow not only for estimating factor effects, but also for judging their reliability and testing their significance.

4.5 Acknowledgements

I would like to acknowledge Jeroen de Groot of TNO Prevention and Health in Leiden, the Netherlands and Virginia B. Kraus of Duke University Medical Center in Durham, NC for providing the guinea pig dataset.

4.6 References

Antti, H., Bollard, M. E., Ebbels, T., Keun, H., Lindon, J. C., Nicholson, J. K., and Holmes, E. (2002) Batch statistical processing of H-1 NMR-derived urinary spectral data, *J.Chemometr.*, **16**, 461-468.

Antti, H., Ebbels, T. M. D., Keun, H. C., Bollard, M. E., Beckonert, O., Lindon, J. C., Nicholson, J. K., and Holmes, E. (2004) 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.*, **73**, 139-149.

Bales, J. R., Higham, D. P., Howe, I., Nicholson, J. K., and Sadler, P. J. (1984) Use of high-resolution proton nuclear magnetic resonance spectroscopy for rapid multi-component analysis of urine, *Clin.Chem.*, **30**, 426-432.

Bendele, A. M. (2001) Animal models of osteoarthritis, *Journal of Musculoskel.Neuron.Interact.*, **1**, 363-376.

Bratchell, N. (1989) Multivaraite response surface modeling by principal component analysis, *J.Chemometr.*, **3**, 579-588.

Cattell, R. B. (1966) The scree test for the number of factors, *Multivariate Behavioral Research*, **1**, 245-276.

Clish, C. B., Davidov, E., Oresic, M., Plasterer, T. N., Lavine, G., Londo, T., Meys, M., Snell, P., Stochaj, W., Adourian, A., Zhang, W., Morel, N., Neumann, E., Verheij, E., Vogels, J. T., Havekes, L. M., Regnier, F., van der Greef, J., and Naylor, S. (2004) Integrative biological analysis of the APOE*3-leiden transgenic mouse *OMICS* 2004, 1:3-13., *Omics*, **1**, 3-13-

Creamer, P. and Hochberg, M. C. (1997) Osteoarthritis, *Lancet*, **350**, 503-508.

Fiehn, O. (2002) Metabolomics-the link between genotypes and phenotypes, *Plant Molecular Biology*, **48**, 155-171.

Gartland, K. P., Sanins, S. M., Nicholson, J. K., Sweatman, B. C., Beddell, C. R., and Lindon, J. C. (1990) Pattern recognition analysis of high resolution 1H NMR spectra of urine. A nonlinear mapping approach to the classification of toxicological data, *NMR Biomed.*, **3**, 166-172.

Gaspari, M., Vogels, J., Wulfert, F., Tas, A. C., Venema, K., Bijlsma, S., Vreeken, R., and van der Greef, J. (2001) Novel strategies in mass spectrometric data handling, *Adv.Mass.Spectrom.*, **15**, 283-296.

Gates, S. C. and Sweeley, Ch. C. (1978) Quantitative metabolic profiling based on gas chromatography, *Clin.Chem.*, **24**, 1663-1673.

Huebner, J. L., Otterness, I. G., Freund, E. M., Caterson, B., and Kraus, V. B. (2001) Collagenase 1 and collagenase 3 expression in a guinea pig model of osteoarthritis, *Arthritis Rheum.*, **41**, 877-890.

Jackson, J., (1991), A User's Guide to Principal Components, Wiley & Sons, New York

Jansen, J. J., Hoefsloot, H. C. J., van der Greef, J., Timmerman, M. E., and Smilde, A. K. (2004) Multilevel Component Analysis of time-resolved metabolomics data, *Anal.Chim.Acta*, **530**, 173-183

Jeffers, J. N. R. (1962) Principal component analysis of designed experiments, *The Statistician*, **12**, 230-242.

Jellum, E. (2001) Chromatography, mass spectrometry and electrophoresis for
diagnosis of human disease, particulary metabolic disorders, in Gehrke, Ch.,
Wixom, R., and Bayer, E. (eds), *Chromatography- a century of discovery 1900-2000*, pp. 270-277.

Keun, H. C., Ebbels, T. M., Bollard, M. E., Beckonert, O., Antti, H., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2004) Geometric trajectory analysis of metabolic responses to toxcicity can define treatment specific profiles, *Chem.Res.Toxicol.*, **17**, 579-587.

Keun, H. C., Ebbels, T. M. D., Antti, H., Bollard, M. E., Beckonert, O., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2003) Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling, *Anal.Chim.Acta*, **490**, 265-276.

Kraus, V. B., Huebner, J. L., Stabler, T., Flahiff, C. M., Setton, L. A., Fink, C., Vilim, V., and Clark, A. G. (2004) Ascorbic acid increases the severity of spontaneous knee osteoarthritis in a guinea pig model, *Arthritis Rheum.*, **50**, 1822-1831.

Lamers, R. A. N., DeGroot, J., Spies-Faber, E. J., Jellema, R. H., Kraus, V. B., Verzijl, N., TeKoppele, J. M., Spijksma, G., Vogels, J. T. W. E., van der Greef, J., and Van Nesselrooij, J. H. J. (2003) Identification of Disease and Nutrient-Related Metabolic Fingerprints in Osteoarthritic Guinea Pigs, *Journal of Nutrition*, **133**, 1776-1780.

Lindon, J. C., Holmes, E., Bollard, M. E., Stanley, E. G., and Nicholson, J. K. (2004) Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis, *Biomarkers*, **9**, 1-31.

Mankin H.J., Dorfman H., Lippiello L, Zarins, A. Biochemical and metabolic abnormalities in articular cartilage from osteo-arthritic human hips. II. Correlation of morphology with biochemical and metabolic data. (1971) *J Bone Joint Surg [Am]*, **53** , 523-37.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979) Multivariate analysis, Academic Press, London

Martens, H. and Martens, M., (2001), Multivariate analysis of quality. An introduction, John Wiley & Sons, Chichester

McAlindon, T. E., Jacques, P., Zhang, Y., Hannan, M. T., Aliabadi, P., Weissman, B., Rush, D., Levy, D., and Felson, D. T. (1996) Do antioxidant micronutrients protect against the development and progression of knee osteoarthritis?, *Arthritis Rheum.*, **39**, 648-656.

Pearce, S. C. and Holland, D. A. (1960) Some applications of multivariate methods in botany, *Applied Statistics*, **9**, 1-7.

Searle, S. R, (1971), Linear models, John Wiley & Sons, New York

Ståhle, L. and Wold, S. (1990) Multivariate analysis of variance (MANOVA), *Chemom.Intell.Lab.Syst.*, **9**, 127-141.

Ten Berge, J. M. F., Kiers, H. A. L., and Van der Stel, V. (1992) Simultaneous component analysis, *Statistica Applicata*, **4**, 277-392.

Timmerman, M. E. Multilevel Component Analysis, *British Journal of Mathematical and Statistical Psychology*; In Press

Van den Brink, P. J. and Ter Braak, C. J. F. (1999) Principal response curves: analysis of time-dependent multivariate responses of biological community to stress, *Environmental Toxicology And Chemistry*, **18**, 138-148.

van der Greef, J., Davidov, E., Verheij, E. R., van der Heijden, R., Adourian,
A. S., Oresic, M., Marple, E. W., and Naylor, S. (2003) The role of metabolomics in Systems Biology, in Harrigan, G. G. and Goodacre, R. (eds), *Metabolic Profiling: Its role in Biomarker Discovery and Gene Function Analysis*, Kluwer Academic Publishers, Boston/Dordrecht/London, pp. 170-198.

van der Greef, J., Tas, A. C., Bouwman, J., Ten Noever de Brauw, M. C., and Schreurs, W. H. P. (1983) Evaluation of field desorption and fast atom bombardment mass spectrometric profiles by pattern recognition techniques, *Anal.Chim.Acta*, **150**, 45-52.

Vogels, J. T. W. E., Tas, A. C., Venekamp, J., and van der Greef, J. (1996) Partial linear fit: a new NMR spectroscopy preprocessing tool for pattern recognition applications, *J.Chemometr.*, **10**, 425-438.

4.7 Appendices

In Appendix 1, the working of ASCA is demonstrated using a simulated dataset.

In Appendix 2, some mathematical properties of the ASCA model are demonstrated. These properties are required for the calculations performed in the manuscript "ANOVA-Simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data". This appendix is built up as follows:

Appendix 2A shows that the column spaces of all submodels are orthogonal to each other. A consequence is that the submodels can be calculated separately and sequentially.

Appendix 2B shows that constraining the scores of the ASCA submodels to a specific subspace is equivalent to projecting the data on this subspace and fitting an unconstrained model to this projected data. The data analysis can then be performed by standard PCA modules.

Appendix 2C shows the orthogonality properties *within* each submodel and between each submodel and its residuals. It shows proof of the percentages of explained variation that are calculated in the manuscript.

The orthogonality properties of the ASCA model as described in Appendix 2 allow the model to be estimated using a relatively simple algorithm. This algorithm is described in Appendix 3.

4.7.1 Appendix 1: Illustration of ASCA using a simulated dataset

In this appendix the use of ASCA is illustrated using a simulated dataset. It is a bivariate example, consisting of two treatment groups, each containing 3 animals. Measurements have been performed on two variables at 5 time-points. The non-systematic variation (e.g. biological variation, measurement error) is constructed to be considerably larger for group 2 than for group 1. The simulated dataset is given in Figure 23. In the figure two groups of points are indicated: group A contains the measurements taken for group 2 at time-point 4 and group B contains the measurements of group 1 at time-point 5.

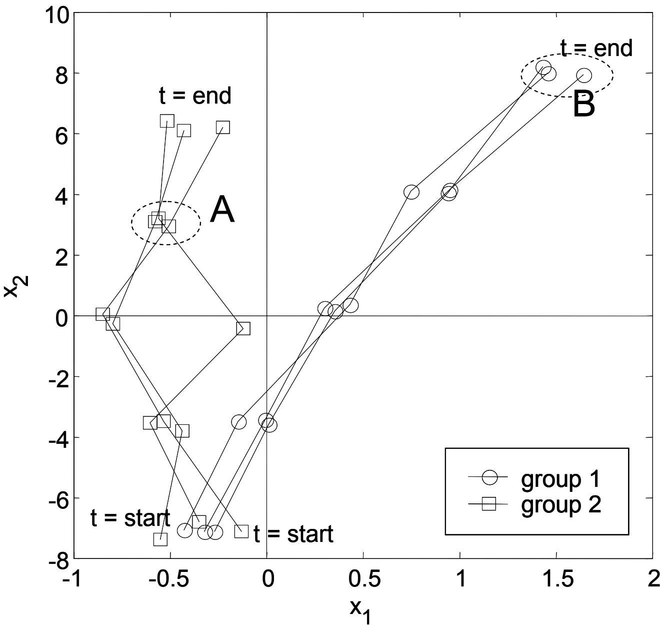

Figure 23 Simulated dataset: the measurements of the 3 individuals at time-point 4 for group 2 are indicated by A and the measurements of the 3 individuals at time-point 5 for group 1 are indicated by B.

This dataset is analyzed using ASCA. As is shown in Appendix 2, the method boils down to a decomposition of the data and a subsequent fitting of SCA models to the obtained matrices. Since this dataset contains only two variables, we will not consider a dimension reduction.

ASCA separates the data in Figure 23 into 3 different contributions: submodel **K**, which is an average trajectory that is equal for all individuals; submodel **K**h that contains the treatment-group specific trajectories as a deviation from submodel **K** and submodel **K**hi$_h$ that contains the individual animal-specific variation as a deviation from submodel **K**h. Since the simulated data is constructed to be mean-centered, the mean is not a part of the model.

The data for submodel **K** can be obtained from the measurements in Figure 23, by calculating the average for all individuals at a specific measurement time-point. The submodel **K** data is given in Figure 24 (indicated by the diamonds), together with the original measurements on both treatment groups (indicated by the open circles and squares). In the figure the dashed arrow indicates the difference between the submodel **K** data at time-point 4 to the average of group A (indicated by a filled square). The dotted arrow indicates the difference between the time-point 5 of submodel **K** to the average of group B (indicated by a filled circle).

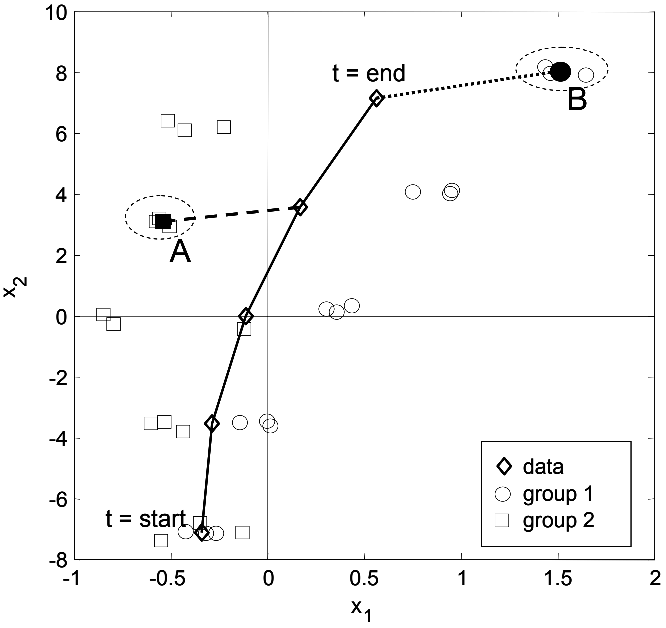

Figure 24 Simulated dataset: Data for submodel K. Again the measurements at time-point 4 for group 2 are indicated by A and the measurements at time-point 5 for group 1 are indicated by B. The filled circle indicates the average of all measurements B and the filled square indicates the average of all measurements A.

The submodel **K** data in Figure 24 is subsequently subtracted from the measurements in Figure 24. The remainder of this subtraction is indicated in Figure 25 by the open circles and squares. The submodel **K**h data can be constructed from this remainder, by calculating the average over all individuals within each dose group, for each time-point (e.g. the submodel **K**h data corresponding to group A is calculated by taking the average of all open squares in Figure 25 belonging to group A). These data points of submodel **K**h have been indicated in Figure 25 by the filled squares and circles. The dashed and dotted arrows from Figure 24 are also indicated in Figure 25 .

Finally, the submodel **K**hi$_h$ data is obtained by subtracting the submodel **K**h data (the filled circles and squares) from the open circles and squares in Figure 25. The remainder

of this subtraction is given in Figure 26. Clearly, no systematic treatment group-related variation is left in the data: submodel **K**hih will only describe information that is specific for each individual animal. The available a priori knowledge concerning the larger non-systematic variation in group 2 is clearly visible in this figure: the spread between the squares is considerably larger than the spread between the circles.

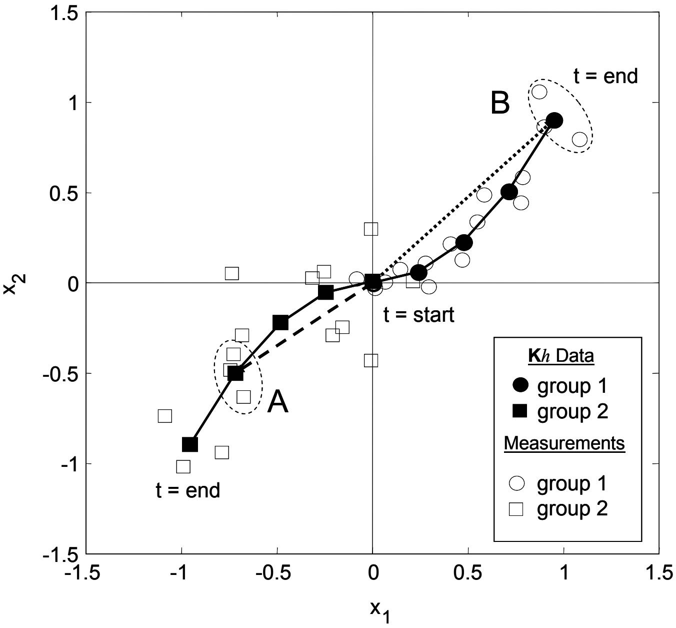

Figure 25 Simulated dataset: Data for submodel Kh. Each filled square and circle now indicates the average of all individuals belonging to either group 1 or 2 at each time-point. A and B indicate the same as in Figure 2, as well as the dotted and the dashed line.

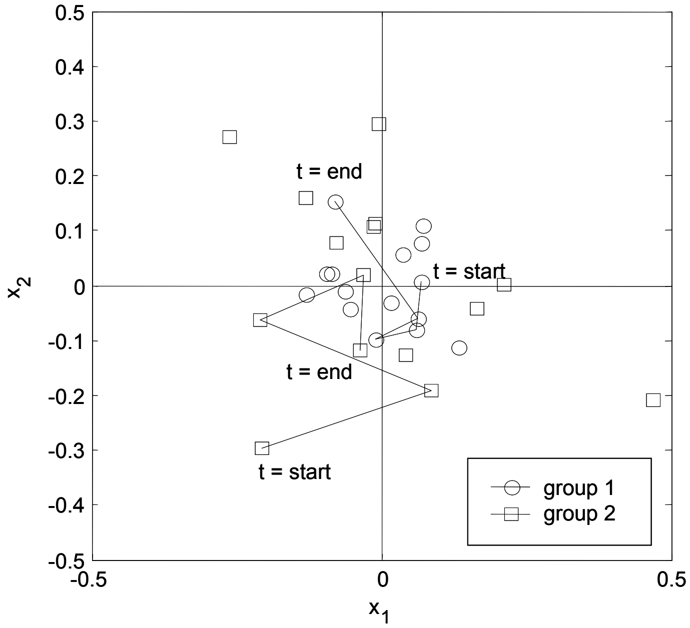

Figure 26 Simulated dataset: Data for submodel Khih. In the figure two time-trajectories, one for each group, are indicated.

4.7.2 Appendix 2: Orthogonality properties of the ASCA model

4.7.2.1 Appendix 2A: Orthogonality between the ASCA submodels

In this appendix the orthogonality of the column spaces of the different submodels is proven. This proof is required for the algorithm used for ASCA. The consequence of this orthogonality is explained after the proof.

The ASCA model can also be written simultaneously for all individuals as:

$$\mathbf{X} = \mathbf{1m}^T + \mathbf{T}_1\mathbf{P}_1^T + \mathbf{T}_2\mathbf{P}_2^T + \mathbf{T}_3\mathbf{P}_3^T + \mathbf{E} \qquad (7)$$

where $\mathbf{X} = \begin{bmatrix} \mathbf{X}_{11_1} \\ \vdots \\ \mathbf{X}_{1I_1} \\ \vdots \\ \mathbf{X}_{H1_H} \\ \vdots \\ \mathbf{X}_{HI_H} \end{bmatrix}$, $\mathbf{T}_1 = \begin{bmatrix} \mathbf{T}_{\mathbf{K}} \\ \vdots \\ \mathbf{T}_{\mathbf{K}} \\ \vdots \\ \mathbf{T}_{\mathbf{K}} \\ \vdots \\ \mathbf{T}_{\mathbf{K}} \end{bmatrix}$, $\mathbf{T}_2 = \begin{bmatrix} \mathbf{T}_{\mathbf{K}1} \\ \vdots \\ \mathbf{T}_{\mathbf{K}1} \\ \vdots \\ \mathbf{T}_{\mathbf{K}H} \\ \vdots \\ \mathbf{T}_{\mathbf{K}H} \end{bmatrix}$ and $\mathbf{T}_3 = \begin{bmatrix} \mathbf{T}_{\mathbf{K}11_1} \\ \vdots \\ \mathbf{T}_{\mathbf{K}1I_1} \\ \vdots \\ \mathbf{T}_{\mathbf{K}H1_H} \\ \vdots \\ \mathbf{T}_{\mathbf{K}HI_H} \end{bmatrix}$ and $\mathbf{E} = \begin{bmatrix} \mathbf{E}_{11_1} \\ \vdots \\ \mathbf{E}_{1I_1} \\ \vdots \\ \mathbf{E}_{H1_H} \\ \vdots \\ \mathbf{E}_{HI_H} \end{bmatrix}$

The constraints that are put on the ASCA-scores (in the general not necessarily balanced case) are:

a. $\mathbf{1}^{\mathbf{T}}\mathbf{T}_{\mathbf{K}} = \mathbf{0}_{R_1}^{\mathbf{T}}$

b. $\sum_{h=1}^{H} I_h \mathbf{T}_{\mathbf{K}h} = \mathbf{0}_{(K \times R_2)}$

c. $\sum_{i_h=1}^{I_h} \mathbf{T}_{\mathbf{K}hi_h} = \mathbf{0}_{(K \times R_3)}$

Using these constraints, the orthogonality between the ASCA-submodels can be proven, as will be done now.

From these constraints it can be proven that matrices $\mathbf{T_1}$, $\mathbf{T_2}$ and $\mathbf{T_3}$ are mean-centered.

$\mathbf{T_1}$: $$\mathbf{1}_{KHI_h}^{\mathbf{T}}\mathbf{T_1} = \left[\mathbf{1}_K^{\mathbf{T}} \ldots \mathbf{1}_K^{\mathbf{T}}\right] \begin{bmatrix} \mathbf{T}_{\mathbf{K}} \\ \vdots \\ \mathbf{T}_{\mathbf{K}} \end{bmatrix} = \sum_{h=1}^{H}\sum_{i_h=1}^{I_h} \mathbf{1}_K^{\mathbf{T}}\mathbf{T}_{\mathbf{K}} = \sum_{h=1}^{H}\sum_{i_h=1}^{I_h} \mathbf{0}_{R_1} = \mathbf{0}_{R_1}$$

$\mathbf{T_2}$: $$\mathbf{1}_{KHI_h}^{\mathbf{T}}\mathbf{T_2} = \mathbf{1}_{KHI_h}^{\mathbf{T}} \begin{bmatrix} \mathbf{T}_{\mathbf{K}1} \\ \vdots \\ \mathbf{T}_{\mathbf{K}1} \\ \vdots \\ \mathbf{T}_{\mathbf{K}H} \\ \vdots \\ \mathbf{T}_{\mathbf{K}H} \end{bmatrix} = \mathbf{1}_K^{\mathbf{T}}\left(\sum_{h=1}^{H} I_h \mathbf{T}_{\mathbf{K}h}\right) = \mathbf{1}_K^{\mathbf{T}}\left(\mathbf{0}_{(K \times R_2)}\right) = \mathbf{0}_{R_2}$$

$$\mathbf{T}_3\text{:}\qquad \mathbf{1}_{KHI_h}^{\mathsf{T}}\mathbf{T}_3=\mathbf{1}_{KHI_h}^{\mathsf{T}}\begin{bmatrix}\mathbf{T}_{\mathbf{K}11_1}\\ \vdots\\ \mathbf{T}_{\mathbf{K}1I_1}\\ \vdots\\ \mathbf{T}_{\mathbf{K}H1_H}\\ \vdots\\ \mathbf{T}_{\mathbf{K}HI_H}\end{bmatrix}=\mathbf{1}_K^{\mathsf{T}}\left(\sum_{h=1}^{H}\sum_{i_h=1}^{I_h}\mathbf{T}_{\mathbf{K}hi_h}\right)=\mathbf{1}_K^{\mathsf{T}}\left(\sum_{h=1}^{H}\mathbf{0}_{(K\times R_3)}\right)=\mathbf{0}_{R_3}$$

From the preceding proofs of mean-centering it follows that all score matrices and therefore all submodels are orthogonal to the matrix $\mathbf{1m}^{\mathsf{T}}$ that contains the column averages.

The orthogonality between the different submodels follows from the orthogonality of their scores:

$$\mathbf{T}_1\textbf{ and }\mathbf{T}_2\text{:}\qquad \mathbf{T}_1^{\mathsf{T}}\mathbf{T}_2=\begin{bmatrix}\mathbf{T}_{\mathbf{K}}\\ \vdots\\ \mathbf{T}_{\mathbf{K}}\\ \vdots\\ \mathbf{T}_{\mathbf{K}}\\ \vdots\\ \mathbf{T}_{\mathbf{K}}\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}\mathbf{T}_{\mathbf{K}1}\\ \vdots\\ \mathbf{T}_{\mathbf{K}1}\\ \vdots\\ \mathbf{T}_{\mathbf{K}H}\\ \vdots\\ \mathbf{T}_{\mathbf{K}H}\end{bmatrix}=\sum_{h=1}^{H}\sum_{i_h=1}^{I_h}\mathbf{T}_{\mathbf{K}}^{\mathsf{T}}\mathbf{T}_{\mathbf{K}h}=\mathbf{T}_{\mathbf{K}}^{\mathsf{T}}\left(\sum_{h=1}^{H}I_h\mathbf{T}_{\mathbf{K}h}\right)=\mathbf{T}_{\mathbf{K}}^{\mathsf{T}}\mathbf{0}_{(K\times R_2)}=\mathbf{0}_{(R_1\times R_2)}$$

T$_1$ and T$_3$:

$$\mathbf{T}_1^{\mathsf{T}}\mathbf{T}_3=\begin{bmatrix}\mathbf{T}_{\mathbf{K}}\\ \vdots\\ \mathbf{T}_{\mathbf{K}}\\ \vdots\\ \mathbf{T}_{\mathbf{K}}\\ \vdots\\ \mathbf{T}_{\mathbf{K}}\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}\mathbf{T}_{\mathbf{K}11}\\ \vdots\\ \mathbf{T}_{\mathbf{K}1I_h}\\ \vdots\\ \mathbf{T}_{\mathbf{K}H1}\\ \vdots\\ \mathbf{T}_{\mathbf{K}HI_H}\end{bmatrix}=\sum_{h=1}^{H}\sum_{i_h=1}^{I_h}\mathbf{T}_{\mathbf{K}}^{\mathsf{T}}\mathbf{T}_{\mathbf{K}hi_h}=\mathbf{T}_{\mathbf{K}}^{\mathsf{T}}\left(\sum_{h=1}^{H}\sum_{i_h=1}^{I_h}\mathbf{T}_{\mathbf{K}hi_h}\right)=\mathbf{T}_{\mathbf{K}}^{\mathsf{T}}\left(\sum_{h=1}^{H}\mathbf{0}_{(K\times R_3)}\right)=\mathbf{0}_{(R_1\times R_3)}$$

T$_2$ and T$_3$:

$$\mathbf{T}_2^{\mathsf{T}}\mathbf{T}_3=\begin{bmatrix}\mathbf{T}_{\mathbf{K}1}\\ \vdots\\ \mathbf{T}_{\mathbf{K}1}\\ \vdots\\ \mathbf{T}_{\mathbf{K}H}\\ \vdots\\ \mathbf{T}_{\mathbf{K}H}\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}\mathbf{T}_{\mathbf{K}11}\\ \vdots\\ \mathbf{T}_{\mathbf{K}1I_h}\\ \vdots\\ \mathbf{T}_{\mathbf{K}H1}\\ \vdots\\ \mathbf{T}_{\mathbf{K}HI_H}\end{bmatrix}=\sum_{h=1}^{H}\sum_{i_h=1}^{I_h}\mathbf{T}_{\mathbf{K}h}^{\mathsf{T}}\mathbf{T}_{\mathbf{K}hi_h}=\sum_{h=1}^{H}I_h\mathbf{T}_{\mathbf{K}h}^{\mathsf{T}}\left(\sum_{i_h=1}^{I_h}\mathbf{T}_{\mathbf{K}hi_h}\right)=\sum_{h=1}^{H}\mathbf{T}_{\mathbf{K}h}^{\mathsf{T}}\mathbf{0}_{(K\times R_3)}=\mathbf{0}_{(R_2\times R_3)}$$

Finally, since the model in equation (1) is estimated using a least-squares minimization, it follows that the residuals **E** are orthogonal to the regressors of the model: i.e. the column mean $\mathbf{m}^{\mathbf{T}}$ and the scores $\mathbf{T_1}$, $\mathbf{T_2}$ and $\mathbf{T_3}$ [1]. This orthogonality will be proven in Appendix 1C.

Consequences:

The consequence of the proof given in this Appendix, is that the submodels can be obtained individually and sequentially [2]. In mathematical terms this means that:

$$\underset{\mathbf{T's,P's}}{\arg\min}\left\|\mathbf{X}-\mathbf{1m}^{\mathbf{T}}-\mathbf{T_1P_1^T}-\mathbf{T_2P_2^T}-\mathbf{T_3P_3^T}\right\|^2=\underset{\mathbf{m}}{\arg\min}\left\|\mathbf{X}-\mathbf{1m}^{\mathbf{T}}\right\|^2+\underset{\mathbf{T_1,P_1}}{\arg\min}\left\|\mathbf{X}-\mathbf{T_1P_1^T}\right\|^2+\underset{\mathbf{T_2,P_2}}{\arg\min}\left\|\mathbf{X}-\mathbf{T_2P_2^T}\right\|^2+$$
$$+\underset{\mathbf{T_3,P_3}}{\arg\min}\left\|\mathbf{X}-\mathbf{T_3P_3^T}\right\|^2$$

Where $\mathbf{T_1}$, $\mathbf{T_2}$ and $\mathbf{T_3}$ are constrained as described before izn this appendix.

4.7.2.2 Appendix 2B: Proof that the least-squares minimization of a constrained model is equal to the least-squares minimization of an unconstrained model of projected data on the space where the constraint is valid.

In this appendix proof is given that the results of a constrained PCA-type model can also be obtained by constraining the data on which the analysis is performed. The consequence of this proof is again given after the proof itself.

J is a projection matrix (**JJ** = **J** and $\mathbf{J}^{\mathbf{T}}$ = **J**), **I** is the identity matrix.

The minimization of a linearly constrained PCA model is given by:

$(\mathbf{T,P})$ = $\underset{\mathbf{T}\in\Re(\mathbf{J}),\mathbf{P}}{\arg\min}\left\|\mathbf{X}\text{ - }\mathbf{TP}^{\mathbf{T}}\right\|^2$ where $\mathbf{P}^{\mathbf{T}}\mathbf{P}=\mathbf{I}$, $\mathbf{T}^{\mathbf{T}}\mathbf{T}$ is a diagonal matrix in which the elements are sorted from large to small and $\mathbf{T}=\mathbf{JT}$.

The minimization of a PCA model on constrained data is given by:

$(\tilde{\mathbf{T}},\tilde{\mathbf{P}})$ = $\underset{\tilde{\mathbf{T}},\tilde{\mathbf{P}}}{\arg\min}\left\|\mathbf{JX}\text{ - }\tilde{\mathbf{T}}\tilde{\mathbf{P}}^{\mathbf{T}}\right\|^2$ where $\tilde{\mathbf{P}}^{\mathbf{T}}\tilde{\mathbf{P}}=\mathbf{I}$ and $\tilde{\mathbf{T}}^{\mathbf{T}}\tilde{\mathbf{T}}$ is a diagonal matrix in which the elements are sorted from large to small.

The question now is: $\left(\tilde{\mathbf{T}},\tilde{\mathbf{P}}\right)\overset{?}{=}\left(\mathbf{T},\mathbf{P}\right)$

$\left\|\mathbf{X}-\mathbf{TP^T}\right\|^2$ can be rewritten as:

$$\left\|\mathbf{X}-\mathbf{TP^T}\right\|^2=\left\|(\mathbf{I}-\mathbf{J})\mathbf{X}\right\|^2+\left\|\mathbf{JX}-\mathbf{TP^T}\right\|^2$$

since:

$$\left\|\mathbf{X}-\mathbf{TP^T}\right\|^2=\left\|(\mathbf{I}-\mathbf{J})\mathbf{X}+\mathbf{JX}-\mathbf{TP^T}\right\|^2=tr\left(\left((\mathbf{I}-\mathbf{J})\mathbf{X}+\mathbf{JX}-\mathbf{TP^T}\right)^\mathbf{T}\left((\mathbf{I}-\mathbf{J})\mathbf{X}+\mathbf{JX}-\mathbf{TP^T}\right)\right)=$$

$$=tr\left(\mathbf{X^T}(\mathbf{I}-\mathbf{J})^\mathbf{T}(\mathbf{I}-\mathbf{J})\mathbf{X}\right)+$$
$$+tr\left(\mathbf{X^T}(\mathbf{I}-\mathbf{J})^\mathbf{T}\mathbf{JX}+\mathbf{X^T}(\mathbf{I}-\mathbf{J})^\mathbf{T}\mathbf{TP^T}+\mathbf{X^T}\mathbf{J^T}(\mathbf{I}-\mathbf{J})\mathbf{X}+\mathbf{X^T}\mathbf{J^T}\mathbf{JX}-\mathbf{X^T}\mathbf{J^T}\mathbf{TP^T}-\mathbf{PT^T}\mathbf{JX}+\mathbf{PT^T}\mathbf{TP^T}\right)$$

where $\mathbf{X^T}(\mathbf{I}-\mathbf{J})^\mathbf{T}\mathbf{TP^T}=\mathbf{X^T}(\mathbf{I}-\mathbf{J})^\mathbf{T}\mathbf{JTP^T}=\mathbf{X^T}\mathbf{0}\mathbf{TP^T}=\mathbf{0}$ since $\mathbf{T}=\mathbf{JT}$.

such that:

$$\left\|\mathbf{X}-\mathbf{TP^T}\right\|^2=tr\left(\mathbf{X^T}(\mathbf{I}-\mathbf{J})^\mathbf{T}(\mathbf{I}-\mathbf{J})\mathbf{X}\right)+tr\left(\mathbf{X^T}\mathbf{J^T}\mathbf{JX}-\mathbf{X^T}\mathbf{J^T}\mathbf{TP^T}-\mathbf{PT^T}\mathbf{JX}+\mathbf{PT^T}\mathbf{TP^T}\right)=$$

$$=\left\|(\mathbf{I}-\mathbf{J})\mathbf{X}\right\|^\mathbf{2}+\left\|\mathbf{JX}-\mathbf{TP^T}\right\|^\mathbf{2}$$

This means that:

$(\mathbf{T},\mathbf{P})$ =

$$\underset{\mathbf{T}\in\Re(\mathbf{J}),\,\mathbf{P}}{\arg\min}\left\|\mathbf{X}-\mathbf{TP^T}\right\|^2=\underset{\mathbf{T}\in\Re(\mathbf{J}),\,\mathbf{P}}{\arg\min}\left(\left\|(\mathbf{I}-\mathbf{J})\mathbf{X}\right\|^2+\left\|\mathbf{JX}-\mathbf{TP^T}\right\|^2\right)=\left\|(\mathbf{I}-\mathbf{J})\mathbf{X}\right\|^2+\underset{\mathbf{T},\mathbf{P}}{\arg\min}\left\|\mathbf{JX}-\mathbf{TP^T}\right\|^2=$$

$$=\underset{\mathbf{T},\mathbf{P}}{\arg\min}\left\|\mathbf{JX}-\mathbf{TP^T}\right\|^2=\underset{\tilde{\mathbf{T}},\tilde{\mathbf{P}}}{\arg\min}\left\|\mathbf{JX}-\tilde{\mathbf{T}}\tilde{\mathbf{P}}^\mathbf{T}\right\|^2=\left(\tilde{\mathbf{T}},\tilde{\mathbf{P}}\right)$$

Note that the constraint that **T** should be in the range of **J** becomes redundant when it is estimated from **JX**.

Since all submodels obtained from the ASCA models are based on centering in certain directions and every such centering is a projection step, every submodel can be estimated by an unconstrained PCA-model on properly centered data.

A similar proof of this equality is given in the paper about CANDELINC by Carrol et al. [3].

Consequences:

The consequence of the proof given in this appendix is, that the following minimization (obtained from Appendix 1A):

$$\underset{\mathbf{T's,P's}}{\arg\min}\left\|\mathbf{X}-\mathbf{1m^T}-\mathbf{T_1P_1^T}-\mathbf{T_2P_2^T}-\mathbf{T_3P_3^T}\right\|^2=\underset{\mathbf{m}}{\arg\min}\left\|\mathbf{X}-\mathbf{1m^T}\right\|^2+\underset{\mathbf{T_1,P_1}}{\arg\min}\left\|\mathbf{X}-\mathbf{T_1P_1^T}\right\|^2+\underset{\mathbf{T_2,P_2}}{\arg\min}\left\|\mathbf{X}-\mathbf{T_2P_2^T}\right\|^2+$$
$$+\underset{\mathbf{T_3,P_3}}{\arg\min}\left\|\mathbf{X}-\mathbf{T_3P_3^T}\right\|^2$$

is equal to minimizing :

$$\underset{\mathbf{m}}{\arg\min}\left\|\mathbf{X}-\mathbf{1m^T}\right\|^2+\underset{\mathbf{T_1,P_1}}{\arg\min}\left\|\mathbf{X_1}-\mathbf{T_1P_1^T}\right\|^2+\underset{\mathbf{T_2,P_2}}{\arg\min}\left\|\mathbf{X_2}-\mathbf{T_2P_2^T}\right\|^2+\underset{\mathbf{T_3,P_3}}{\arg\min}\left\|\mathbf{X_3}-\mathbf{T_3P_3^T}\right\|^2$$

where:

$$\mathbf{X_1}=\begin{bmatrix}\overline{\mathbf{X}}_{\mathbf{c}}\\ \vdots\\ \overline{\mathbf{X}}_{\mathbf{c}}\\ \vdots\\ \overline{\mathbf{X}}_{\mathbf{c}}\\ \vdots\\ \overline{\mathbf{X}}_{\mathbf{c}}\end{bmatrix},\quad \mathbf{X_2}=\begin{bmatrix}\overline{\mathbf{X}}_{\mathbf{d}1}\\ \vdots\\ \overline{\mathbf{X}}_{\mathbf{d}1}\\ \vdots\\ \overline{\mathbf{X}}_{\mathbf{d}H}\\ \vdots\\ \overline{\mathbf{X}}_{\mathbf{d}H}\end{bmatrix}\text{ and }\mathbf{X_3}=\begin{bmatrix}\mathbf{X}_{\mathbf{I}11}\\ \vdots\\ \mathbf{X}_{\mathbf{I}1I_1}\\ \vdots\\ \mathbf{X}_{\mathbf{I}H1}\\ \vdots\\ \mathbf{X}_{\mathbf{I}HI_H}\end{bmatrix}$$

Where:

$$\overline{\mathbf{X}}_{\mathbf{C}}=HI_h^{-1}\sum_{h=1}^{H}\sum_{i_h=1}^{I_h}\mathbf{X}_{\mathbf{C}hi_h}\quad\text{and}\quad\mathbf{X}_{\mathbf{C}}=\begin{bmatrix}\mathbf{X}_{\mathbf{C}11_1}\\ \vdots\\ \mathbf{X}_{\mathbf{C}HI_h}\end{bmatrix}=\mathbf{X}\text{ - }\mathbf{1m^T}$$

$$\overline{\mathbf{X}}_{\mathbf{D}h}=I_h^{-1}\sum_{i_h=1}^{I_h}\mathbf{X}_{\mathbf{D}hi_h}\quad\text{and}\quad\mathbf{X}_{\mathbf{D}hi_h}=\mathbf{X}_{\mathbf{C}hi_h}-\overline{\mathbf{X}}_{\mathbf{C}}$$

and $\mathbf{X}_{\mathbf{I}hi_h}=\mathbf{X}_{\mathbf{D}hi_h}-\overline{\mathbf{X}}_{\mathbf{D}h}$

Such that the constraints on $\mathbf{X_1}$, $\mathbf{X_2}$ and $\mathbf{X_3}$ are given by:

$\mathbf{X_1}$: $\quad\mathbf{1^T}\overline{\mathbf{X}}_c=\mathbf{0}_J^{\mathbf{T}}$

$\mathbf{X_2}$: $\quad\sum_{h=1}^{H}I_h\overline{\mathbf{X}}_{\mathbf{d}h}=\mathbf{0}_{(K\times J)}$

$\mathbf{X_3}$: $\quad\sum_{i_h=1}^{I_h}\mathbf{X}_{\mathbf{I}hi_h}=\mathbf{0}_{(K\times J)}$

Using a proof that is analogous to Appendix 2A, it can be proven that:

$$\mathbf{X_1^TX_2}=\mathbf{X_1^TX_3}=\mathbf{X_2^TX_3}=\mathbf{1m^TX_1}=\mathbf{1m^TX_2}=\mathbf{1m^TX_3}=\mathbf{0}$$

such that:

$$\|\mathbf{X}\|^2 = \|\mathbf{1m^T}\|^2 + \|\mathbf{X_1}\|^2 + \|\mathbf{X_2}\|^2 + \|\mathbf{X_3}\|^2$$

4.7.2.3 Appendix 2C: Orthogonality within the ASCA submodels

The percentage of explained variation of each of the submodels **K**, **K**h and **K**hi_h can also be calculated for partitions of matrices $\mathbf{X_1}$, $\mathbf{X_2}$ and $\mathbf{X_3}$. Proof of this is given below. The consequences of this proof are given after the proof itself.

The total variation in the data **X** can be split up into different matrices, according to equation (2).

$$\|\mathbf{X}\|^2 = \|\mathbf{1m^T}\|^2 + \|\mathbf{X_1}\|^2 + \|\mathbf{X_2}\|^2 + \|\mathbf{X_3}\|^2 \qquad (8)$$

where:

$$\mathbf{X_1} = \begin{bmatrix} \overline{\mathbf{X}}_{\mathbf{c}} \\ \vdots \\ \overline{\mathbf{X}}_{\mathbf{c}} \\ \vdots \\ \overline{\mathbf{X}}_{\mathbf{c}} \\ \vdots \\ \overline{\mathbf{X}}_{\mathbf{c}} \end{bmatrix}, \quad \mathbf{X_2} = \begin{bmatrix} \overline{\mathbf{X}}_{\mathbf{d}1} \\ \vdots \\ \overline{\mathbf{X}}_{\mathbf{d}1} \\ \vdots \\ \overline{\mathbf{X}}_{\mathbf{d}H} \\ \vdots \\ \overline{\mathbf{X}}_{\mathbf{d}H} \end{bmatrix} \text{ and } \mathbf{X_3} = \begin{bmatrix} \mathbf{X}_{\mathbf{I}11} \\ \vdots \\ \mathbf{X}_{\mathbf{I}1I_1} \\ \vdots \\ \mathbf{X}_{\mathbf{I}H1} \\ \vdots \\ \mathbf{X}_{\mathbf{I}HI_H} \end{bmatrix}$$

Submodel **K**

Submodel **K** is given by:

$$\mathbf{X_1} = \mathbf{T}_1\mathbf{P}_1^{\mathrm{T}} + \mathbf{E}_1 \qquad (9)$$

Because of the least squares estimation the following properties hold: $\mathbf{T_1^TE_1 = 0}$ and $\mathbf{E_1P_1 = 0}$. From this it follows that :

$$\|\mathbf{X_1}\|^2 = \|\mathbf{T_1P_1^T}\|^2 + \|\mathbf{E_1}\|^2 \qquad (10)$$

Furthermore, since $\mathbf{P_1^TP_1} = \mathbf{I}$:

$$\|\mathbf{X_1}\|^2 = \left\|\mathbf{t}_\mathbf{1}^\mathbf{1}\mathbf{p}_\mathbf{1}^{\mathbf{1}\,\mathbf{T}}\right\|^2 + \ldots + \left\|\mathbf{t}_\mathbf{1}^{R_1}\mathbf{p}_\mathbf{1}^{R_1\,\mathbf{T}}\right\|^2 + \|\mathbf{E_1}\|^2 \tag{11}$$

Where $\mathbf{t}_\mathbf{1}^{r_1}\mathbf{p}_\mathbf{1}^{r_1\,\mathbf{T}}$ is the r1'th component of submodel **K**

Thus the variation explained by submodel **K** can also be calculated per component.

Submodel **K***h*

Submodel **K***h* is given by:

$$\mathbf{X_2 = T_2P_2^T + E_2} \tag{12}$$

Because of the least squares estimation the following properties hold: $\mathbf{T_2^TE_2 = 0}$ and $\mathbf{E_2P_2 = 0}$. From this it follows that:

$$\|\mathbf{X_2}\|^2 = \left\|\mathbf{T_2P_2}^\mathbf{T}\right\|^2 + \|\mathbf{E_2}\|^2 \tag{13}$$

Furthermore, since $\mathbf{P_2}^\mathbf{T}\mathbf{P_2} = \mathbf{I}$:

$$\|\mathbf{X_2}\|^2 = \left\|\mathbf{t}_2^\mathbf{1}\mathbf{p}_\mathbf{2}^{\mathbf{1}\,\mathbf{T}}\right\|^2 + \ldots + \left\|\mathbf{t}_\mathbf{2}^{R_1}\mathbf{p}_\mathbf{2}^{R_1\,\mathbf{T}}\right\|^2 + \|\mathbf{E_2}\|^2 \tag{14}$$

Where $\mathbf{t}_2^{r_2}\mathbf{p}_\mathbf{2}^{r_2\,\mathbf{T}}$ is the r_2'th component of submodel **K***h*

Thus the variation explained by submodel **K***h* can also be calculated per component.

Submodel **K***hi*$_h$

Submodel **K***hi*$_h$ is given by:

$$\mathbf{X_3 = T_3P_3^T + E_3} \tag{15}$$

From the least squares properties $\mathbf{T_3}^{\mathbf{T}}\mathbf{E_3} = \mathbf{0}$ and $\mathbf{E_3P_3} = \mathbf{0}$ it follows that:

$$\left\|\mathbf{X_3}\right\|^2 = \left\|\mathbf{T_3P_3}^{\mathbf{T}}\right\|^2 + \left\|\mathbf{E_3}\right\|^2 \tag{16}$$

Furthermore, since $\mathbf{P_3}^{\mathbf{T}}\mathbf{P_3} = \mathbf{I}$:

$$\left\|\mathbf{X}_3\right\|^2 = \left\|\mathbf{t_3^1p_3^1}^{\mathbf{T}}\right\|^2 + \ldots + \left\|\mathbf{t_3}^{R_1}\mathbf{p_3}^{R_1\,\mathbf{T}}\right\|^2 + \left\|\mathbf{E_3}\right\|^2 \tag{17}$$

Where $\mathbf{t}_3^{r_3}\mathbf{p_3}^{r_3\,\mathbf{T}}$ is the r_3'th component of submodel $\mathbf{K}hi_h$

Thus the variation explained by submodel $\mathbf{K}hi_h$ can also be calculated per component.

Consequences:

A consequence of the proof in Appendix 1C is that matrix **E** in equation (1) of Appendix 1A can be written as:

$\mathbf{E} = \mathbf{E_1} + \mathbf{E_2} + \mathbf{E_3}$

The orthogonality properties of these residuals follow from the known properties of PCA, they are demonstrated for submodel 1 (but also hold for submodels 2 and 3):

$\mathbf{T_1^T E_1} = \mathbf{E_1 P_1} = \mathbf{0}$

It is known from Appendix 1B that $\mathbf{X_1^T X}_2 = \mathbf{0}$

Together with the fact that $\mathbf{T}_1 \in \Re(\mathbf{X}_1)$ and $\mathbf{E}_2 \in \Re(\mathbf{X}_2)$ it follows that:

$\mathbf{T_1^T X_2} = \mathbf{T_1^T E_2} = \mathbf{0}$,

using the analogous proof for submodels 2 and 3 it follows that:

$\mathbf{T_1^T E} = \mathbf{T_2^T E} = \mathbf{T_3^T E} = \mathbf{0}$

and that the column space of the residuals of the ASCA model is orthogonal to the column spaces of all three ASCA submodels, such that:

$$\|\mathbf{X}\|^2 = \|\mathbf{1m^T} + \mathbf{T_1P_1^T} + \mathbf{T_2P_2^T} + \mathbf{T_3P_3^T} + \mathbf{E}\|^2 = \|\mathbf{1m^T}\|^2 + \|\mathbf{T_1P_1^T}\|^2 + \|\mathbf{T_2P_2^T}\|^2 + \|\mathbf{T_3P_3^T}\|^2 + \|\mathbf{E}\|^2$$

Due to this equality the percentage of explained variation by the entire ASCA model can be calculated as follows:

$$\exp_{\mathbf{tot}} = \frac{\|\mathbf{T_1P_1}^{\mathbf{T}}\|^2 + \|\mathbf{T_2P_2}^{\mathbf{T}}\|^2 + \|\mathbf{T_3P_3}^{\mathbf{T}}\|^2}{\|\mathbf{X\text{ - }1m^T}\|^2} \times 100\%$$

4.7.3 Appendix 3: the Algorithm

First we present the algorithm for balanced data, such that $I_1 = \ldots = I_H$

1. The dataset **X** has dimensions (HI_hK x J)
2. Mean center the data: $\mathbf{X_C} = \mathbf{X} - \mathbf{1m^T}$
3. $\mathbf{X_C}$ contains HI$_h$ data blocks $\mathbf{X}_{\mathbf{C}hi_h}$ of dimensions (K x J), containing the data corresponding to each individual
4. Calculate the 'mean data' $\overline{\mathbf{X}}_{\mathbf{C}}$ over all animals hi_h = 11_1 ... HI_H

: $$\overline{\mathbf{X}}_{\mathbf{C}} = \frac{\sum_{h=1}^{H}\sum_{i_h=1}^{I_h}\mathbf{X}_{\mathbf{C}hi_h}}{HI_h}$$

5. **submodel K:** perform a PCA on $\overline{\mathbf{X}}_{\mathbf{C}}$
6. subtract $\overline{\mathbf{X}}_{\mathbf{C}}$ from all $\mathbf{X}_{\mathbf{C}11} \ldots \mathbf{X}_{\mathbf{C}HI_H}$ to obtain $\mathbf{X}_{\mathbf{D}11} \ldots \mathbf{X}_{\mathbf{D}HI_H}$
7. Calculate the 'mean data' $\overline{\mathbf{X}}_{\mathbf{D}h}$ for each dose group $h = 1 \ldots H$:

$$\overline{\mathbf{X}}_{\mathbf{D}h} = \frac{\sum_{i_h=1}^{I_h}\mathbf{X}_{\mathbf{D}hi_h}}{I_h}$$

8. **submodel Kh:** perform a PCA on $\begin{bmatrix}\overline{\mathbf{X}}_{\mathbf{D}1} \\ \vdots \\ \overline{\mathbf{X}}_{\mathbf{D}H}\end{bmatrix}$
9. subtract $\overline{\mathbf{X}}_{\mathbf{D}h}$ from all $\mathbf{X}_{\mathbf{D}h1} \ldots \mathbf{X}_{\mathbf{D}hI_H}$ to obtain $\mathbf{X}_{\mathbf{I}11} \ldots \mathbf{X}_{\mathbf{I}HI_H}$

10. **submodel K**hi_h: perform a PCA on $\begin{bmatrix} \mathbf{X}_{\mathbf{I}11} \\ \vdots \\ \mathbf{X}_{\mathbf{I}HI_H} \end{bmatrix}$

When the data is unbalanced (i.e. contains a different number of animals I_h for different dose groups), this can be implemented into the algorithm in step 8. This step then changes into:

8a. **submodel K**h: perform a PCA on $\begin{bmatrix} \sqrt{I_1}\,\overline{\mathbf{X}}_{\mathbf{D}1} \\ \vdots \\ \sqrt{I_H}\,\overline{\mathbf{X}}_{\mathbf{D}H} \end{bmatrix}$

8b. The scores $\mathbf{T}_{\mathbf{K}h}$ obtained from 8a then have to be re-weighted by $\sqrt{I_h}^{-1}$

4.7.4 References

1. Draper, N.R. and H. Smith, *Applied Regression Analysis*. Wiley series in probability and statistics, ed. W.A. Shewhart and S.S. Wilks. 1998, New York: John Wiley & sons. 137.
2. Timmerman, M.E., Multilevel Component Analysis. *British Journal of Mathematical and Statistical Psychology*; In Press
3. Carroll, J.D., S. Pruzansky, and J.B. Kruskal, *CANDELINC: A general approach to multidimensional analysis of many-way data arrays with linear constraints on parameters*, in *Psychometrika*. 1980. p. 3.

5 Visualising homeostatic capacity: hepatoxicity in the rat

5.1 Introduction

In Life Sciences, the development in recent years of novel technologies to analyze biological systems has been impressive and has opened up new strategies for studying biology. The focus on systems biology using the so-called '-omics' technologies in combination with advanced bio-statistical and bioinformatics methods has created a significant step towards a better understanding of biology; specifically in understanding critical remaining challenging issues as discussed in a recent issue of Nature Biotechnology (1).

Studies in mammalian Systems Biology have pointed out that the analysis of biological systems in view of the coverage of important biological levels such as transcripts, proteins and metabolites still needs to be further technologically enhanced. However, at the current state of these technologies it is already possible to reveal unique and significant new levels of information from biological systems (2-4). Such observations specifically emphasize typical system characteristics: for example the connectivity and interdependence within biological systems, potential non-linearity and the emerging properties at different levels of complexity of these systems (5-7). More applied research in systems biology focuses either on systems pathology to discover biomarkers of disease or on systems pharmacology or toxicology to reveal the response of a system towards an exogenous intervention (8).

The changes in biology when moving from a healthy towards a disease state are based on a disturbance of the self-organization principles of an organism (9). An observed loss of homeostasis is often only quantitatively reflected in biomarker patterns in a later stage of the system response (10), while earlier stages of this response are often characterized by a change in pathway dynamics to compensate for the exogenous perturbation (as described in the dynamic disease concept (11)). These dynamics (determined by the system properties of the organism) determine whether and how fast the perturbed organism regains homeostasis. This is referred to as the ‘homeostatic capacity’ (12). A system-wide analysis on these dynamics gives a more complete view on such a system response than a view on a relatively small number of biomarkers.

Studying these complex dynamic systems and describing them in a model is extremely difficult. However, when specific questions are asked and optimized experiments are available, impressive results have been reported (e.g. in studying aspects of metabolic syndrome (13)).

For the diagnostic measurement of the homeostatic capacity of an organism a perturbation of the system followed by measurement of its dynamic response profile, especially at an early disease stage, is a good approach (8). Of specific interest in this response is the ‘homeostatic capacity’: the ability of a biological system to retain its homeostasis despite the presence of an exogenous perturbation. An example of this is the oral glucose tolerance test, which is used to detect the deregulation of glucose metabolism in early states of diabetes type 2. This experiment is typically performed in the most simplified way: perturbing the system by drinking a glucose solution and measuring the resulting glucose level after 2-3 hours to evaluate the homeostatic capacity. These measurements can be extended with concentrations of insulin or glucagon, which are directly involved in the regulatory system for blood glucose concentration.

When time-resolved ‘omics’-technologies are used, rather than the measurement of individual biomarkers, more details about the behavior of the system become available and the perturbation of the metabolism can be measured at an earlier stage, as mentioned before. For example in toxicological research often insight is needed in the time-dependent response of a system to evaluate the different stages of the impact of a toxic compound on an organism. These dose-response studies are preferably performed using a robust and relatively uncomplicated analytical strategy, while trying to obtain as much information about the system under investigation as possible. Nuclear Magnetic Resonance (NMR) (14) and Mass spectrometry (15) are often used techniques in metabolic fingerprinting. Studies have shown that a toxic insult, specifically by compounds having an effect on organs that are important in metabolism, can be monitored using time-resolved metabolic fingerprinting on urine (16-19).

In this paper, the toxicity of Bromobenzene (BB) in the rat is investigated. BB is a model hepatoxicant that causes necrosis in the liver and in the kidneys. The hepatic transformation and the toxicity of BB in the rat have been reported in detail (20-25). Earlier research has shown that the reaction of metabolism to BB can be monitored using time-resolved metabolic fingerprinting on urine and that the reaction of the metabolism to BB toxicity can be differentiated based on animal species (26).

The focus in this study is on the dose-dependence of the reaction of metabolism to a toxic insult with BB. Therefore rats are divided into different groups, to which different doses of BB are administered. The urinary metabolite composition of the rats is monitored at different points in time using NMR. The concentration of a metabolite might be dependent on the factors imposed by the experimental design, as well as influenced by the concentration of other, related metabolites. These relationships are determined from the available data using an analysis of the collected metabolic fingerprinting data.

Univariate data collected from an experimental design is conventionally analyzed using ANOVA (27, 28), while for the analysis of the variation in multivariate data usually Principal Component Analysis (PCA) is used (29, 30). However, important parts of information contained within multivariate data from an experimental design are lost when either of these methods are used for the analysis (31). Therefore ANOVA-SCA (or ASCA) is chosen for the analysis of this dataset (32, 33). The ASCA model disentangles the different contributions to the variation caused by the different factors in the experimental design and takes the covariance between the urinary concentrations of different metabolites into account as well. The response of the urine composition to the toxic insult is evaluated using the ASCA model, specifically with respect to the homeostatic capacity of the rat for bromobenzene.

5.2 Materials and Methods

5.2.1 Samples and data acquisition

The experiment is conducted on 45 rats that are randomly divided into 5 different groups, each containing 9 rats. Three of these groups receive a dose of toxicant at the beginning of the experiment: the low, medium and high dose groups respectively receive 0.5, 2.0 and 5.0 mmol BB / kg body weight, dissolved in corn oil. The 2 remaining groups are the 'vehicle group' that receives only corn oil and the control group that receives no treatment. During the experiment the rats received water *ad libitum* but no food. The focus of the study is the homeostatic capacity of the rats to the administration of this toxicant.

After 6 hours urine is obtained from 3 of the rats in each group, which are subsequently sacrificed. The livers of the sacrificed rats are collected for visual inspection. This procedure is repeated after 24 and 48 hours in the experiment. The data contains one missing value: after 48 hours the urine of one of the 3 rats in the high dose group could not be obtained. The samples collected in time reflect the ability of the rats to regain homeostasis of their metabolism.

Animals were kept under controlled conditions, and the welfare of the animals was maintained in accordance with the general principles governing the use of animals in toxicity experiments of the European Communities (Directive 86/609/EEC) and Dutch legislation (The Experiments on Animals Act, 1997).

A description of the sampling and subsequent NMR analysis is given by Heijne *et al.* (34). The obtained NMR spectra were vector normalized to compensate for differences in dilution of the urine.

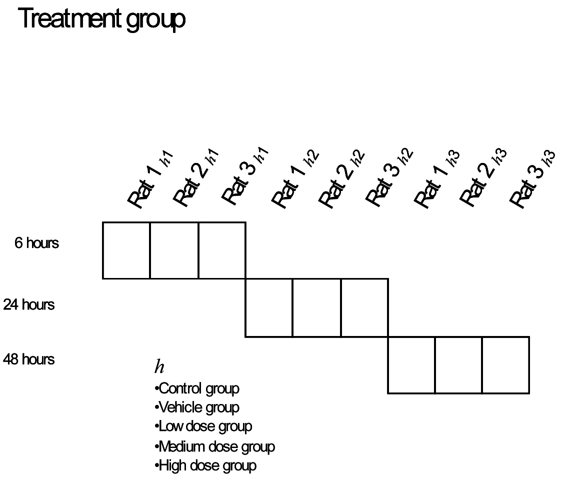

Figure 27 Design of the bromobenzene experiment. Each square corresponds to a sample in the design and therefore to an NMR spectrum. This figure represents the design of one treatment group. Note that the design of the total experiment consists of five of these groups of squares.

5.2.2 Experimental design and the variation in the data

The described dataset contains 5 treatment groups ($h = 1(\textbf{control}),\ldots,5(\textbf{high dose})$). Measurements are performed at 3 time-points during the study ($k = 1(\textbf{6 hours}),\ldots,3(\textbf{48 hours})$): the measurement time-points are crossed with treatment groups. Measurements are performed on 3 rats *per* treatment group *per* measurement time-point ($i_{hk} = 1\ldots3$, but $i_{\textbf{high},48} = 1\ldots2$): this means that the rats are both nested within treatment group and within measurement time-point. The design of the BB experiment is given in Figure 27. The NMR spectrum of each sample contains information on 310 NMR channels ($j = 1\ldots310$) that are identical for all samples: each NMR channel corresponds to one chemical shift in the spectrum.

The total variation in the data can be split up into different contributions. These contributions are schematically depicted for one simulated metabolite in Figure 28. The variation consists of a time-course that is equal for all rats, since all rats receive a perturbation with the same toxin and are subject to identical experimental circumstances.

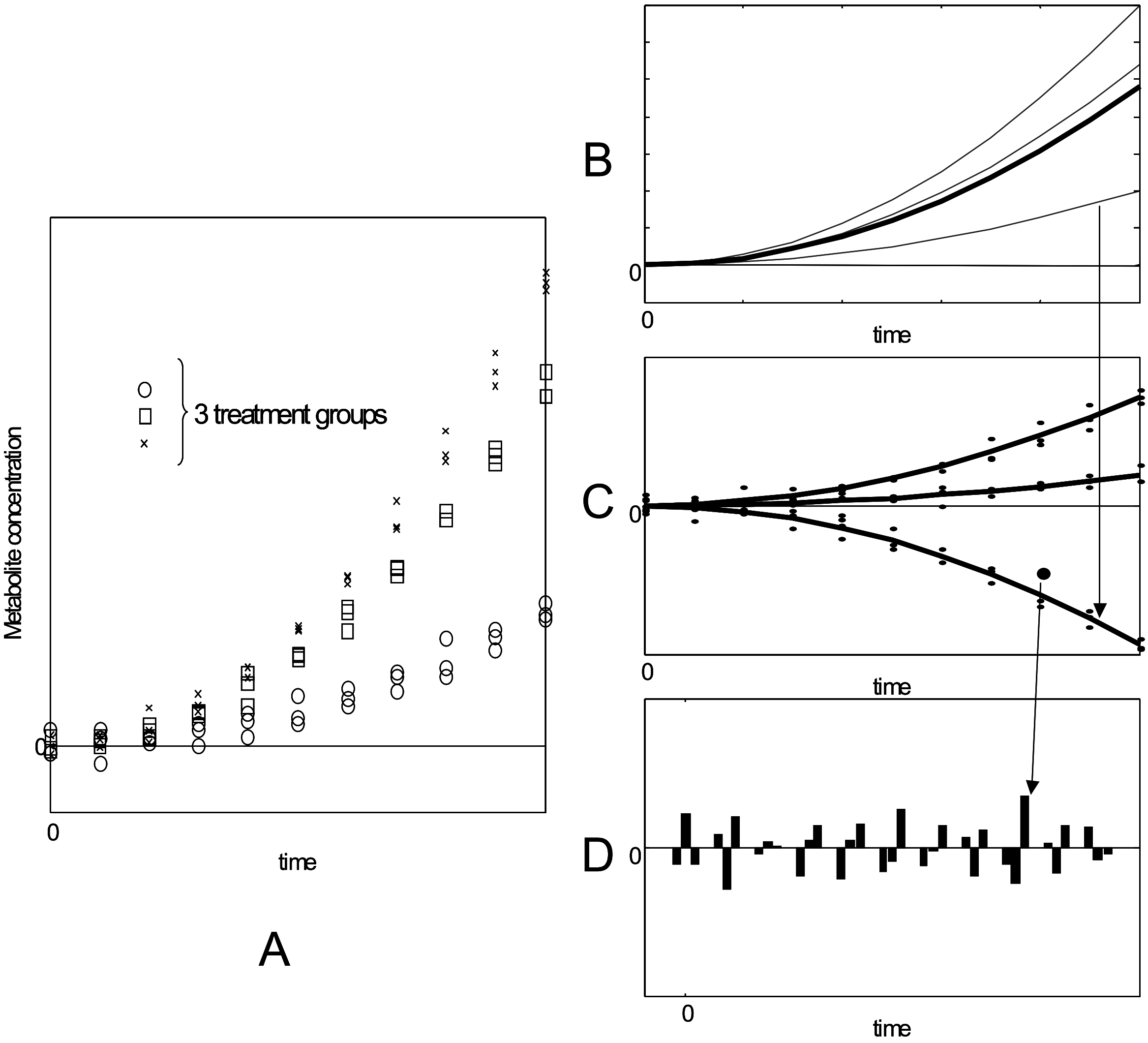

Figure 28 Univariate schematic representation of the variation in the BB dataset. A. All variation in the dataset, this variation can be decomposed into the contributions B: time-course equal for all rats, C: treatment group specific time-course, D: individual rat specific variation. In B the treatment-group specific profiles are indicated by the thinner lines, in C the individual-specific related profiles are indicated by the dots.

Rats are divided into different treatment groups, which will lead to an additional specific time-course of the metabolism that is related to the administered dose of BB. Furthermore, due to biological variability each rat will have an individual contribution to the variation. Information about the homeostatic capacity of the rats to BB is described by the treatment-group specific variation as described by **C** in Figure 28: when a specific treatment group is close to the control group, it is in homeostasis. When this is not the case, the treatment group has not (yet) regained homeostasis. For multivariate data the representation of the variation in Figure 28 is severely oversimplified, because the different variables (i.e. the different NMR channels in this metabolic fingerprinting dataset) covary. The urinary concentrations of different metabolites and therefore the sizes of different peaks in each spectrum will be interrelated. This covariance is generally not identical for the different contributions defined in Figure 28. The relative importance of metabolites in each contribution is crucial to elucidating the biological phenomena underlying the information in a metabolic fingerprinting dataset: the covariance between the metabolites is important to obtain a system-wide view on the metabolic mechanisms underlying the homeostatic capacity of the rats for BB.

5.2.3 Data Analysis - ASCA

The basis of ASCA is an ANOVA equation. ANOVA separates the observed variation in the data into different contributions, corresponding to different factors and interactions in an experimental design. An appropriate ANOVA equation for the BB study, that reflects the split-up of the variation into the contributions defined in Figure 28, is given by equation [1].

$$[1] \qquad x_{i_{hk}j} = \mu_j + \alpha_{kj} + (\alpha\beta)_{hkj} + (\alpha\beta\gamma)_{i_{hk}j}$$

where $x_{i_{hk}j}$ indicates the data value of rat i_{hk} (nested within dose group h and measurement time-point k) on NMR channel j; μ_j is the average value of the values $x_{i_{hk}j}$ for each channel j; α_{kj} is the factor 'time' of channel j and varies between time-points k; $(\alpha\beta)_{hkj}$ signifies the treatment group specific variation of channel j (it is not a multiplication) and $(\alpha\beta\gamma)_{i_{hk}j}$ indicates the variation of channel j that is specific for each individual rat. Equation [1] differs only from a 'standard' ANOVA equation in the presence of the additional variable index j.

The available data is used to estimate the model parameters in equation [1]. These estimates are given in terms of the data in equation [2].

[2] $$x_{i_{hk}j} = x_{...j} + \left(x_{..kj} - x_{...j}\right) + \left(x_{h.kj} - x_{..kj}\right) + \left(x_{i_{hk}j} - x_{h.kj}\right)$$

where $x_{...j}$ is the average of the values in the data for each variable j, $x_{..kj}$ is the average of the values in the data for each time-point k and variable j and $x_{h.kj}$ is the average of all values in the data for each time-point k, treatment group h and variable j.

All values of $x_{i_{hk}j}$ can be placed into a data matrix **X**, such that each row is an NMR spectrum of one sample (all values j of one rat i_{hk}). This can be done for the different factors and interactions in the design as well, such that matrices of size equal to **X** are obtained, which leads to the matrix-wise expression of equation [2] given in equation [3].

[3] $$\mathbf{X} = \mathbf{1m}^{\mathbf{T}} + \mathbf{X_a} + \mathbf{X_{(ab)}} + \mathbf{X_{(abg)}}$$

where **1** is a column vector containing ones, **m** is a size J column vector containing the values of $x_{...j}$, matrix $\mathbf{X_a}$ contains all values of $\left(x_{..kj} - x_{...j}\right)$, $\mathbf{X_{(ab)}}$ of $\left(x_{h.kj} - x_{..kj}\right)$ and $\mathbf{X_{(abg)}}$ of $\left(x_{i_{hk}j} - x_{h.kj}\right)$.

The variation in $\mathbf{X_a}$, $\mathbf{X_{(ab)}}$ and $\mathbf{X_{(abg)}}$ can be modeled using Simultaneous Component Analysis (SCA) models. SCA is a generalization of PCA for fitting a model to multiple matrices simultaneously (35-37). SCA (like PCA) uses the covariance between the different NMR channels to obtain a simplified view on the variation that is better interpretable than the original data. By combining the SCA models of the different matrices, the ASCA model in equation [4] is obtained.

[4] $$\mathbf{X} = \mathbf{1m}^{\mathbf{T}} + \mathbf{T_a P_a^T} + \mathbf{T_{(ab)} P_{(ab)}^T} + \mathbf{T_{(abg)} P_{(abg)}^T} + \mathbf{E}$$

where $\mathbf{T}_a$ denotes the score matrix and $\mathbf{P}_a$ the loading matrix of the SCA model of $\mathbf{X}_a$, $\mathbf{T}_{(ab)}$ and $\mathbf{P}_{(ab)}$ of $\mathbf{X}_{(ab)}$ and matrices $\mathbf{T}_{(abg)}$ and $\mathbf{P}_{(abg)}$ of $\mathbf{X}_{(abg)}$ respectively; all loading matrices in the ASCA model are constrained as $\mathbf{P}^T\mathbf{P} = \mathbf{I}$; $\mathbf{E}$ is a matrix of the same dimensions as $\mathbf{X}$ that contains the residuals of the ASCA model.

The three different submodels in equation [4] will be subsequently indicated as follows: $\mathbf{T}_a\mathbf{P}_a^T$ will be called the ‘time’ submodel, $\mathbf{T}_{(ab)}\mathbf{P}_{(ab)}^T$ will be called the ‘interaction time x treatment’ submodel and $\mathbf{T}_{(abg)}\mathbf{P}_{(abg)}^T$ will be called the ‘Individual rat‘ submodel.

The information about each NMR channel (and therefore about the urinary metabolites) is captured by the loadings of the submodels. Interpretation of these loadings can reveal the biological background of the phenomena observed in the scores. An algorithm in MATLAB that fits the ASCA model to data is available online at:

http://www-its.chem.uva.nl/research/pac/people/phdstudents/jeroen_jansen.html

5.3 Results and Discussion

5.3.1 Contributions to the variation

The relative magnitude of each contribution can be determined from the estimates in equation [2] (31, 32). The variation of the factor ‘time’ contributes 36 % to the total variation in the dataset and the treatment groups contribute 40 % to the total variation in the data (that is described by the ‘interaction time x treatment’ submodel). The individual rats contribute the remaining 24 % to the total variation. This means that the treatment group-specific variation, that contains information about the homeostatic capacity, is the largest contributor to the variation. The information about the homeostatic capacity of the rats to BB can be obtained from the scores of the ‘interaction time x treatment’ submodel.

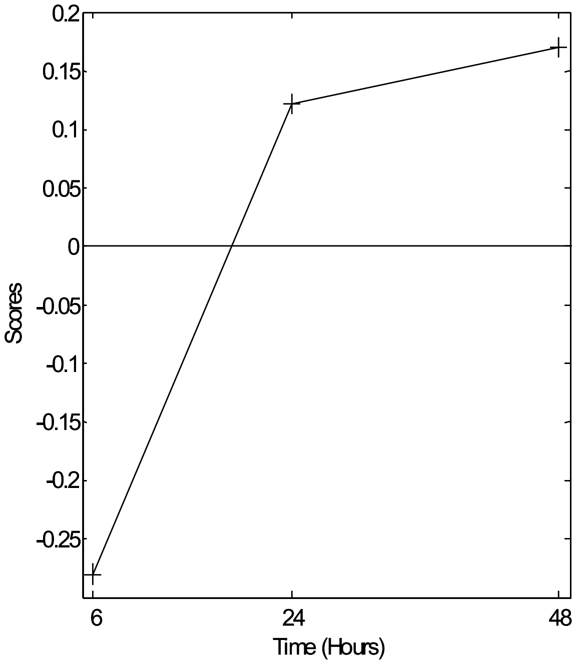

Figure 29 Scores of the 'time' submodel

5.3.2 ‘Time’ submodel

The ‘time’ submodel, containing one component, describes 88% of the average dynamic variation of all animals in the data. This submodel therefore gives a good view on the dynamic variation that is equal for all rats in the experiment. The scores of the ‘time’ submodel are given in Figure 29. These scores clearly show that the change in the composition of the urine is largest between the measurement time-points at 6 and 24 hours. The loadings of the ‘time’ submodel are given in Figure 30 **I**, the loadings of the other submodels are given in Figure 30 **II** and **III**. Of each loading, the ten largest peaks are indicated in the figure by their chemical shift. The metabolites that are annotated to these chemical shifts are given in Table 2.

In Figure 30 I, the peaks corresponding to allantoine, glycerol and creatinine are positive and the peaks corresponding to citric, hippuric and acetic acid are negative. Furthermore, the scores in Figure 29 increase in time. The combination of this information shows that a toxic insult by BB is characterized by a decreasing urinary concentration of citric, hippuric and acetic acid and an increasing concentration of allantoine, glycerol and creatinine.

Comparison of the loadings in Figure 30 I to the loadings in II and III shows that for the 10 largest peaks, allantoine is identified as unique to the 'time' submodel. This means that its concentration varies equally in time for all treatment groups.

Table 2 Chemical shifts and annotated urinary metabolites that are important in the ASCA loadings

Chemical Shift (ppm)	Compound
Allantoine	5.38
hippuric acid	3.9675
Glycerol	3.7525
Glycerol	3.675
Creatinine	3.0475
citric acid	2.735 2.6975 2.5825 2.5425
acetic acid	2.055
Trimethylamine N-oxide (TMAO)	3.285
Trimethylamine	2.93
Dimethylamine	2.91
N-acetyl-glycoproteins	2.075

5.3.3 'Interaction time x treatment' submodel

The first component of the 'interaction time x treatment' submodel explains 71 % of the variation in the data: the submodel gives a view on most of the treatment group-specific variation. The scores of this submodel are given in Figure 31.

First of all, the figure shows that the scores of the vehicle and control groups are almost equal. The effect that corn oil has on the metabolism of the rats is negligible in comparison with BB and therefore in the remainder both groups will be collectively referred to as 'control'.

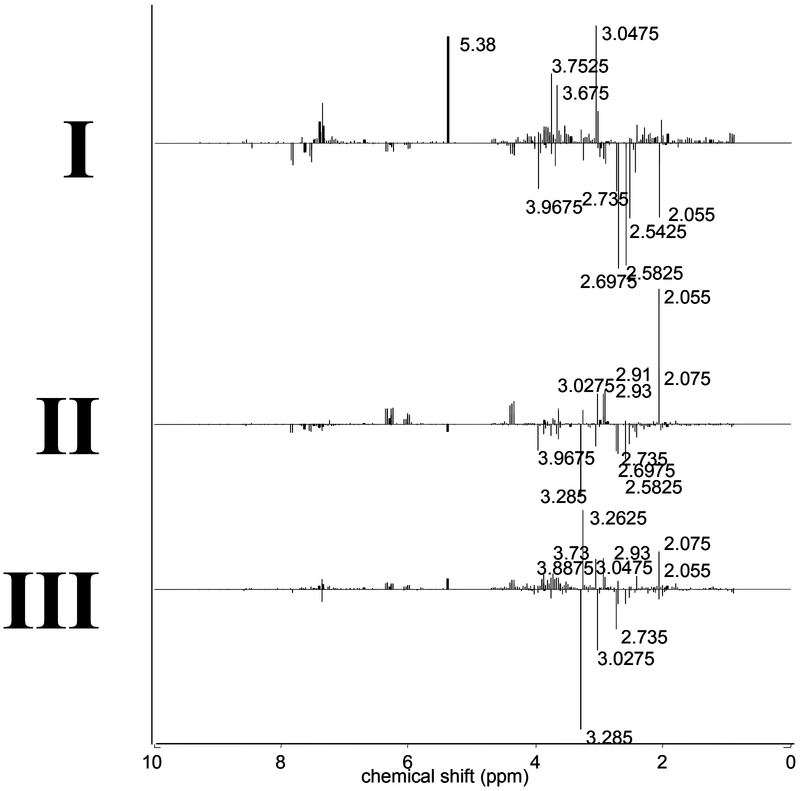

Figure 30 Loadings of the three submodels: I. Time submodel, II. Interaction time x treatment submodel, III. Individual rat submodel

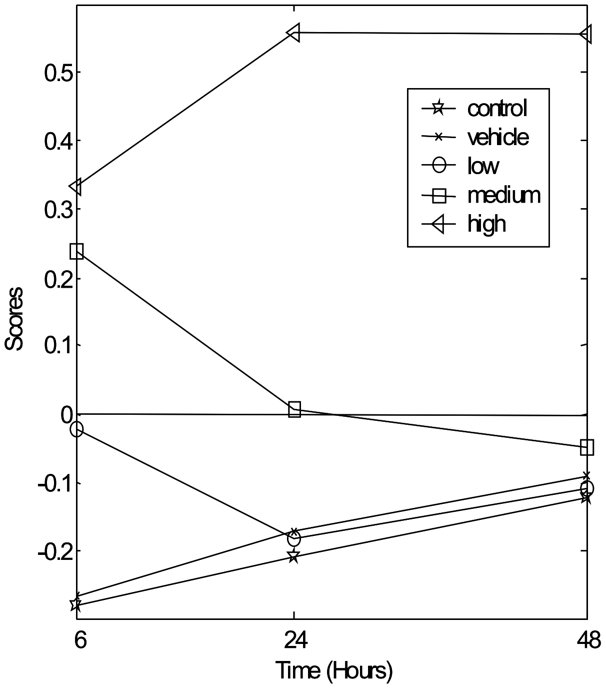

Figure 31 Scores of the 'interaction time x treatment' submodel

The scores in Figure 31 indicate the homeostatic capacity of the rats for BB. Clearly the low and medium dose groups regain homeostasis within the time-span of the experiment: their scores have returned to the range of the control scores at the end of the experiment. The high dose group does not regain homeostasis after 48 hours: a dose of 5.0 mg/kg body weight is too large for the homeostatic capacity of the rats. Figure 31 shows that also for the low and medium dose groups the response of the metabolism to the toxic insult is quantitative: it takes longer for the medium dose group to regain homeostasis than for the low dose group.

The visual inspection of the rat livers shows that no effects of BB are visible in the livers of the low dose group; the livers of the medium dose group animals exhibit very slight BB-related effects after 24 hours that almost disappear after 48 hours. The rats in the high dose group show significantly more severe effects in their livers (centrilobular necrosis) after 24 hours that do not disappear after 48 hours. This means that these clinical measurements corroborate with the results of the metabolic fingerprinting.

Figure 31 shows that a larger effect of BB on the metabolism results in a higher score on the ‘Interaction time x treatment’ submodel. Comparison of these scores to the loadings of this submodel in Figure 30 **II** shows that larger doses of BB induce an increased urinary concentration of acetic acid, N-acetyl glycoproteins, dimethylamine, trimethylamine and malate, as well as a decreased concentration of citric acid, TMAO and hippuric acid.

5.3.4 ‘Individual rat’ submodel

The score values of the ‘individual rat’ submodel are mainly of interest for statistical and diagnostic purposes. Therefore they are not shown and only very briefly discussed. They show that the spread between the scores of the low and medium dosed rats is comparable to that of the control: the variation between the rats within these dose groups is comparable in size to the normal biological variation. The high dose rats have an increased spread: higher doses of BB induce an increased variation between the animals within the treatment group.

The loadings of the ‘individual rat’ submodel are given in Figure 30 **III**. The largest peak in the loadings corresponds to TMAO. This means that the spread in urinary concentration of TMAO is increased between the rats of the high dose group. Since TMAO is also identified in the ‘interaction treatment x time’ submodel, it also varies between the treatment groups. Hippuric acid, which is large in both the loadings of the ‘time’ and of the ‘interaction time x treatment’ submodel and absent from the ‘individual rat’ submodel, shows variation induced by BB that is similar for all rats within a treatment group.

5.4 Conclusions

This study shows that the homeostatic capacity of an organism for a toxicant can be determined by the described approach. This approach consists of a combination of a perturbation of the metabolism of the organism, monitoring the time-resolved response of the metabolism by metabolic fingerprinting and finally visualizing this response using ASCA. It enables monitoring of the response of the metabolism, as well as the identification of metabolites causing this response. Information about these metabolites can be used further to increase the mechanistic insight behind the perturbation and response.

This systems biology-based approach is very general: the combination of an insult of a biological system, monitoring its response and the analysis of this response using a dedicated multivariate data analysis technique (like ASCA) is widely applicable for determining specific system properties in all fields of ‘omics’, as well as in many other fields of research.

5.5 Acknowledgements

I wish to acknowledge Jack T.W.E. Vogels for the biological interpretation of the model results. Robert-Jan A.N. Lamers, Wilbert H. Heijne and Joop H. van Nesselrooij are acknowledged for making available the bromobenzene data.

5.6 References

1. Nature Biotechnology **22(10)**.

2. Clish, C. B., Davidov, E., Oresic, M., Plasterer, T. N., Lavine, G., Londo, T., Meys, M., Snell, P., Stochaj, W., Adourian, A., Zhang, X., Morel, N., Neumann, E., Verheij, E. R., Vogels, J. T. W. E., Havekes, L. M., Afeyan, N., Regnier, F., van der Greef, J. & Naylor, S. (2004) OMICS **8,** 3-13.

3. Davidov, E., Clish, C. B., Oresic, M., Meys, M., Stochaj, W., Snell, P., Lavine, G., Londo, T., Adourian, A., Zhang, W., Johnston, M., Morel, N., Marple, E. W., Plasterer, T. N., Neumann, E., Verheij, E. R., Vogels, J. T. W. E., Havekes, L. M., Van der Greef, J. & Naylor, S. (2004) OMICS **8,** 267-288.

4. Oresic, M., Clish, C. B., Davidov, E., Verheij, E. R., Vogels, J. T. W. E., Havekes, L. M., Neumann, E., Adourian, A., Naylor, S., Van der Greef, J. & Plasterer, T. N. (2004) Applied Bioinformatics **3,** 205-217.

5. Sheldrake, R. (1991) The Rebirth of Nature (Bantam Books, NJ).

6. Capra, F. (1996) The Web of Life (Anchor Books, New York).

7. Laszlo, E. (1996) The Systems View of the World: A Holistic View of Our Time (Hampton Press, New Jersey).

8. Van der Greef, J. & McBurney, R. Submitted.

9. Oltvai, Z. N. & Barabasi, A. L. (2002) Science **298,** 763-764.

10. Van der Greef, J., Stroobant, P. & van der Heijden, R. (2004) Curr. Opin. Chem. Biol. **8,** 559-565.

11. Glass, L. & Mackey, M. C. (1988) From Clocks to Chaos: "The Rythms of Life" (Princeton University Press, Princeton).

12. Frank, R. & Tankersley, C. (2002) Environmental Health Perspectives **110**.

13. Kitano, H., Oda, K., Kimura, T., Matsuoka, Y., Csete, M., Doyle, J. & Muramatsu, M. (2004) Diabetes **53,** S6-15.

14. Nicholson, J. K., Connelly, J. C., Lindon, J. C. & Holmes, E. (2002) in Nature Reviews Drug Discovery, Vol. 1, pp. 153.

15. van der Greef, J., Davidov, E., Verheij, E. R., van der Heijden, R., Adourian, A. S., Oresic, M., Marple, E. W., Naylor, S., Harrigan, G. G. & Goodacre, R. (2003) in Metabolic Profiling: Its role in Biomarker Discovery and Gene Function Analysis (Kluwer Academic Publishers, Boston/Dordrecht/London), pp. 170.

16. Lindon, J. C., Nicholson, J. K., Holmes, E., Antti, H., Bollard, M. E., Keun, H., Beckonert, O., Ebbels, T. M., Reily, M. D., Robertson, D. G., Stevens, G. J., Luke, P., Breau, A. P., Cantor, G. H., Bible, R. H., Niederhauser, U., Senn, H., Schlotterbeck, G., Sidelmann, U. G., Laursen, S. M., Tymiak, A., Car, B. D., Lehman-McKeeman, L., Colet, J. M., Loukaci, A. & Thomas, C. (2003) in Toxicology and Applied Pharmacology, Vol. 187, pp. 137.

17. Kleno, T. G., Kiehr, B., Baunsgaard, D. & Sidelmann, U. G. (2004) Biomarkers **9,** 116.

18. Beckwith-Hall, B. M., Nicholson, J. K., Nicholls, A. W., Foxall, P. J. D., Lindon, J. C., Connor, S. C., Abdi, M. & Holmes, E. (1998), Vol. 11, pp. 260.

19. Azmi, J., Griffin, J. L., Antti, H., Shore, R. F., Johansson, E., Nicholson, J. K. & Holmes, E. (2002) in The Analyst, Vol. 127, pp. 271.

20. Den Besten, C., Brouwer, A., Rietjens, I. M. & Van Bladeren, P. J. (1994) in Human and Experimental Toxicology, Vol. 13, pp. 866.

21. Miller, N. E., Thomas, D. & Billings, R. E. (1990) in Drug Metabolism and Disposition, Vol. 18, pp. 304.

22. Monks, T. J., Lau, S. S. & Gillette, J. R. (1982) in Life Science, Vol. 30, pp. 841.

23. Lau, S. S. & Monks, T. J. (1988) in Life Science, Vol. 42, pp. 1259.

24. Casini, A. F., Pompella, A. & Comporti, M. (1985) in American Journal of Pathology, Vol. 118, pp. 225.

25. Thor, H., Svensson, S. A., Hartzell, P. & Orrenius, S. (1981), Vol. 136, pp. 287.

26. Keun, H. C., Ebbels, T. M., Bollard, M. E., Beckonert, O., Antti, H., Holmes, E., Lindon, J. C. & Nicholson, J. K. (2004) **17,** 579-587.

27. Sokal, R. R. & Rohlf, F. J. (1995) Biometry (W.H. Freeman and company, San Francisco).

28. Searle, S. R. (1971) Linear Models (John Wiley & Sons, Inc., New York).

29. Jackson, J. E. (1991) A User's Guide to Principal Components (Wiley-Interscience, New York).

30. Jolliffe, I. T. (2002) Principal Component Analysis (Springer-Verlag, New York).

31. Jansen, J. J., Hoefsloot, H. C. J., van der Greef, J., Timmerman, M. E. & Smilde, A. K. (2005) Analytica Chimica Acta **530,** 173-183.

32. Smilde, A. K., Jansen, J. J., Hoefsloot, H. C. J., van der Greef, J. & Timmerman, M. E. (2005) Bioinformatics**,** In Press.

33. Jansen, J. J., Hoefsloot, H. C. J., Van der Greef, J., Timmerman, M. E. & Smilde, A. K. (Submitted).

34. Heijne, W. H., Lamers, R. J. A. N., Van Bladeren, P. J., Groten, J. P., Van Nesselrooij, J. H. J. & Van Ommen, B.

35. Kiers, H. A. L. & Ten Berge, J. M. F. (1994) British Journal of Mathematical and Statistical Psychology **47,** 109-126.

36. Ten Berge, J. M. F., Kiers, H. A. L. & Van der Stel, V. (1992) Statistica Applicata **4,** 377-392.

37. Timmerman, M. E. & Kiers, H. A. L. (2003) Psychometrika **86,** 105-122.

6 ASCA: analysis of multivariate data obtained from an experimental design

6.1 Introduction

In many designed experiments multivariate data is generated. For example, datasets from functional genomics (metabolomics, proteomics, transcriptomics) contain information about a relatively small number samples measured on a large number of variables. Different sources of variation are present in such datasets. In the analysis of this type of data the design of the experiment as well as the relationship between the different variables should be taken into account: both are interesting to understand the system underlying the variation in the data.

Analysis of Variance (ANOVA) is generally used to analyze data from an experimental design (2, 3). It is a univariate method and therefore it cannot take the covariance between different variables into account. Principal Component Analysis (PCA) or Simultaneous Component Analysis (SCA) that can be seen as PCA for multiple matrices (4), is a widely used method that models the relationships between the different variables in a multivariate dataset by analyzing its covariance or correlation matrix (5, 6). However, SCA does not take the experimental design into account, which means that the different contributions to the variation caused by the experimental design are confounded in the model. This seriously hampers the interpretation of the variation in the data (7). Clearly both ANOVA and SCA only give a limited view.

The idea of separating different types of variation in multivariate data analysis is not new. It originated in botany and pomology (8, 9). Several methods for the analysis of multivariate datasets with an experimental design are available. The classical extension of ANOVA for multivariate data is Multivariate-ANOVA (MANOVA) (10). However, when the number of variables exceeds the number of measured samples, MANOVA breaks down because it cannot handle singular covariance matrices (11). A proposed solution for this breakdown is using Partial Least Squares regression (APLSR, as proposed by Martens (12)), but the precise implementation of ANOVA into this method (i.e. whether the variables encoding the experimental design should be used as regressors or regressands) remains unclear (11, 12). Another multivariate generalization of ANOVA is PC-ANOVA. In this method, first a PCA model is constructed of a multivariate dataset and subsequently the fitted principal components are analyzed using ANOVA (13). This approach has been criticized, since the separate ANOVA's on the different principal components are not mutually independent. This is because in the constructed PCA model the different contributions to the variation by the experimental design are confounded (5).

Another model that is often employed for the analysis of multivariate data from an experimental design is the Structural Equation Model (14). This model is capable of analyzing the experimental design, as well as the relationship between the different variables. However, a Structural Equation Model is difficult to identify when the data contains many variables: strict assumptions about the statistical distribution of the data are required (multivariate normality) and large sample sizes are necessary, which are often not available.

A novel approach for the analysis of multivariate data from a designed experiment is ANOVA – Simultaneous Component Analysis (ASCA). In this method the parameter estimation aspect of ANOVA is merged with PCA, such that the previously mentioned drawbacks of both methods are removed. Thereby a data analysis method is obtained that takes both the covariance between the multiple variables *and* the design of the experiment into account. ASCA has been applied to data analysis problems in psychology (15) and metabolomics (1, 7, 16). In some of these papers a special case of ASCA is used: *Multilevel* Simultaneous Component Analysis (MSCA), which is ASCA for nested designs.

This paper focuses on the framework of the ASCA method, its mathematical properties and its similarities and differences with SCA and ANOVA. First ASCA is explained and subsequently the mathematical properties of ASCA are derived. Several case studies from metabolomics are presented and the relation of ASCA to some other methods is elucidated. Finally some possible extensions of the ASCA model are given.

6.2 Theory

6.2.1 Analysis of Variance

Analysis of Variance (ANOVA) is a widely used technique for univariate data and is one of the basic tools in many fields of research (2, 3). It is used to determine the effect of different experimental factors on the variation in a dataset. Apart from testing hypotheses, ANOVA can be used as a method for parameter estimation, such that the effects of the different levels of each factor can be quantified.

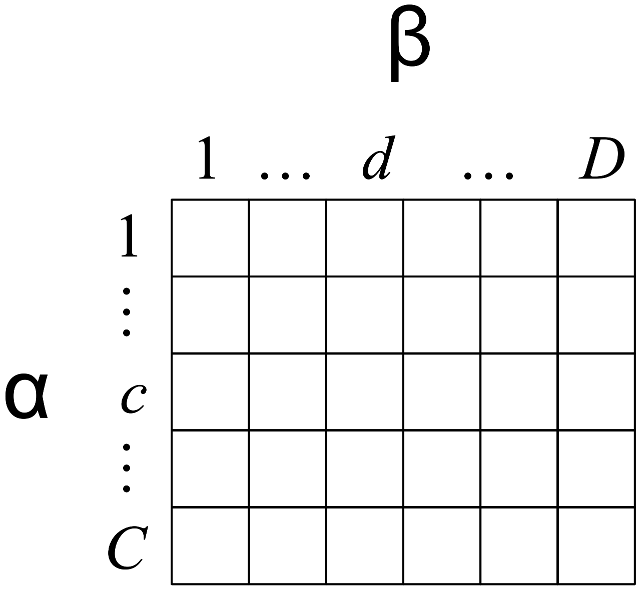

Figure 32 Experimental design where a two-way ANOVA without replication is relevant. Each combination of levels of factor α and factor β contains one measurement. *1…c…C* **is the index for the levels of factor α and** *1…d…D* **the index for levels of factor β**

An example of an ANOVA equation is the 'two-way analysis of variance without replication'. In a design where this ANOVA is relevant, two experimental factors α and β are varied over different levels and at each combination one sample is measured on a dependent variable. Such a design is schematically depicted in Figure 32. An ANOVA equation for this design is given in equation [1].

[1] $x_{cd} = \mu + \alpha_c + \beta_d + \varepsilon_{cd}$

where x_{cd} is the data value observed for the sample on level c and d, μ is an offset term, α_c is the model parameter for factor α on level c, β_d is the parameter of the ANOVA model for factor β on level d and ε_{cd} is the error; $1...c...C$ is the index for the levels of factor α and $1...d...D$ the index for levels of factor β.

The ANOVA equation [1] does not have a unique solution, hence different (equally valid) solutions can be obtained (2). Each solution will have a different set of estimates of parameters μ, α_c and β_d. Which of these solutions is chosen is immaterial for the construction of an ANOVA table.

Table 3 Usual Constraints for the two-way ANOVA model without replication; Substituting all estimates into equation [1] leads to the equality $x_{cd} = x_{cd}$. The value $x_{..}$ is the overall average of variable x; $x_{c.}$ is the average value of all measurements of variable x corresponding to level c; $x_{.d}$ is the average value of all measurements of variable x corresponding to level d.

ANOVA factor	Constraints	Estimate
μ	-	$x_{..}$
α_c	$\sum_{c=1}^{C} \alpha_c = 0$	$x_{c.} - x_{..}$
β_d	$\sum_{d=1}^{D} \beta_d = 0$	$x_{.d} - x_{..}$
$(\alpha\beta)_{cd}$	$\sum_{c=1}^{C} (\alpha\beta)_{cd} = 0 \forall d$ and $\sum_{d=1}^{D} (\alpha\beta)_{cd} = 0 \forall c$	$x_{cd} - x_{c.} - x_{.d} + x_{..}$

To obtain a unique solution to the ANOVA model, constraints have to be imposed on the parameters in equation [1]. The commonly used constraints are referred to by Searle as the 'usual constraints' to the solution (2). For the two-way ANOVA model without replication for the design in Figure 32, the constraints are given in Table 3. The specific constraints in Table 3 give the solution properties that are favorable for the ASCA model. These will be discussed later. The solution of the parameters in the ANOVA model that satisfies the usual constraints is also given in Table 3.

6.2.2 Principal Component Analysis and Simultaneous Component Analysis

6.2.2.1 Principal Component Analysis

Multivariate datasets can be represented as a matrix **X**. Each row of this matrix contains the measurements of one sample n and each column contains the measurement of one variable j. A Principal Component Analysis (PCA) model can then approximate the information in **X** (5, 6). This is done for a matrix $\mathbf{X}_1$ of dimensions $(N_1 \times J)$ in equation [2], where N_1 is the total number of samples in $\mathbf{X}_1$ and J is the total number of variables in matrix $\mathbf{X}_1$.

[2] $$\mathbf{X}_1 = \mathbf{T}_1\mathbf{P}_1^{\mathbf{T}} + \mathbf{E}_1$$

where $\mathbf{T}_1$ is the $(N_1 \times R_1)$ matrix containing the component scores of the model of $\mathbf{X}_1$ and $\mathbf{P}_1$ is the $(J \times R_1)$ matrix containing the loadings; matrix $\mathbf{E}_1$ is the $(N_1 \times J)$ matrix containing the residuals of the model and R_1 is the number of components that is selected for the PCA model of $\mathbf{X}_1$

6.2.2.2 Simultaneous Component Analysis

Quite often, multiple matrices $\mathbf{X}_q$ are available that each contain N_q samples on which the same variables J have been measured (index 1…q...Q relates to different sets of samples). Then two approaches can be chosen for making a component model of these matrices. One approach consists of fitting separate PCA models for each matrix $\mathbf{X}_q$, each containing a score matrix $\mathbf{T}_q$ and a loading matrix $\mathbf{P}_q$, analogous to equation [2]. A serious drawback of using this approach is that the scores in matrices $\mathbf{T}_q$ are not comparable between different q's, as the loading matrix and thereby the basis on which the scores of each object is expressed, differs between objects. This hampers the interpretation of the model.

Alternatively the Simultaneous Component Analysis (SCA) approach can be chosen, where one loading matrix is fitted for all matrices $\mathbf{X}_q$, such that equation [3] is obtained (4, 17, 18).

$$[3] \quad \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_Q \end{bmatrix} = \begin{bmatrix} \mathbf{T}_1 \\ \vdots \\ \mathbf{T}_Q \end{bmatrix} \mathbf{P}^{\mathbf{T}} + \begin{bmatrix} \mathbf{E}_1 \\ \vdots \\ \mathbf{E}_Q \end{bmatrix} = \mathbf{TP}^{\mathbf{T}} + \begin{bmatrix} \mathbf{E}_1 \\ \vdots \\ \mathbf{E}_Q \end{bmatrix}$$

where each matrix $\mathbf{X}_q$ has dimensions $(N_q \times J)$, each matrix $\mathbf{T}_q$ has dimensions $(N_q \times R)$ and R is the number of components fitted for the SCA model; $\mathbf{P}$ has dimensions $(J \times R)$ and $\mathbf{E}_q$ has dimensions equal to $\mathbf{X}_q$; $\mathbf{T}$ is defined here as the vertical concatenation of all matrices $\mathbf{T}_1 \ldots \mathbf{T}_Q$.

When no additional constraints are placed on the component scores $\mathbf{T}_q$, the SCA model in equation [3] is equivalent to PCA on matrix $\begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_Q \end{bmatrix}$ and is estimated using least-squares fitting.

If only one matrix $\mathbf{X}_1$ is analyzed, SCA in equation [3] and PCA in equation [2] are identical. For simplicity, the component model that is used in ASCA will be consistently referred to as SCA.

6.2.2.3 ANOVA-Simultaneous Component Analysis (ASCA)

ASCA is explained here using the design in Figure 32. However, it can be used for the analysis of data with any experimental design. In the description of ANOVA the measurements in the design given in Figure 32 were assumed to be univariate. However, often multivariate measurements are available from a designed experiment. Then each square in Figure 32 represents a measurement on multiple variables indexed as $1 \ldots j \ldots J$. Each data value in a multivariate measurement in this experiment can be represented as x_{cdj}. Analogous to equation [1], the measurements can then be decomposed using the ANOVA equation [4].

$$x_{cdj} = \mu_j + \alpha_{cj} + \beta_{dj} + (\alpha\beta)_{cdj} \qquad [4]$$

where equation [4] represents a series of J ANOVAs. In this equation the error in the ANOVA equation is indicated by $(\alpha\beta)_{cdj}$ instead of ε_{cdj}, which would be logical from equation [1]. This change in notation is chosen to avoid confusion with the error term of the subsequently described ASCA model.

The terms in equation [4] can be estimated using Table 3 (with an additional subscript j for the variable index). Since the values x_{cdj} are obtained as multivariate measurements, they can be collected into a matrix **X** of dimensions $(N \times J)$; where N is equal to the total number of samples collected in the experiment. In Figure 32, N is equal to CD and also to $\sum_{q=1}^{Q} N_q$. Similarly, all estimates of the ANOVA parameters on the right hand side of equation [4] can be collected into matrices. Thereby equation [5] is obtained.

$$\mathbf{X} = \mathbf{1}\mathbf{m}^{\mathbf{T}} + \mathbf{X}_{\mathbf{a}} + \mathbf{X}_{\mathbf{b}} + \mathbf{X}_{(\mathbf{ab})} \qquad [5]$$

where **1** is a size N column vector, $\mathbf{m}^{\mathbf{T}}$ is a size J row vector containing all estimates of μ_j ; matrices $\mathbf{X}_{\mathbf{a}}$, $\mathbf{X}_{\mathbf{b}}$ and $\mathbf{X}_{(\mathbf{ab})}$ contain the estimates of parameters α_{cj}, β_{dj} and $(\alpha\beta)_{cdj}$ respectively.

The rows of matrices $\mathbf{X}_{\mathbf{a}}$ and $\mathbf{X}_{\mathbf{b}}$ are highly structured. All rows related to one level c of factor α are equal in $\mathbf{X}_{\mathbf{a}}$ and analogously all rows of $\mathbf{X}_{\mathbf{b}}$ are equal for each level d of factor β. The decomposition given in equation [5] can be performed using any constraints on the solution. However the ‘usual constraints’ and therefore the estimates in Table 3 make the column spaces of the matrices of the decomposition mutually orthogonal. In that case equation [6] holds in addition to equation [5] (for details, see Appendix 1).

[6] A $$\|\mathbf{X}\|^2 = \|\mathbf{1m}^\mathrm{T}\|^2 + \|\mathbf{X_a}\|^2 + \|\mathbf{X_b}\|^2 + \|\mathbf{X_{(ab)}}\|^2$$

where $\|\mathbf{X}\|^2$ denotes the sum-of-squares of the elements in **X**. Equation [6] shows that by imposing the constraints in Table 3, the variation in **X** is split into independent parts. Equation [6] can be used to determine the contribution of each factor and interaction to the total variation in the data.

SCA component models can be used to approximate the information in matrices $\mathbf{X_a},\mathbf{X_b},\mathbf{X_{(ab)}}$ in equation [5]. The ASCA model corresponding to the ANOVA model given in equation [1] is given by equation [7].

[7] $$\mathbf{X} = \mathbf{1m}^\mathrm{T} + \mathbf{T_a P_a^T} + \mathbf{T_b P_b^T} + \mathbf{T_{(ab)} P_{(ab)}^T} + \mathbf{E}$$

where the SCA component scores of each submodel are given by the matrices indicated by $\mathbf{T_a},\mathbf{T_b},\mathbf{T_{(ab)}}$ and the submodel loadings are given by matrices $\mathbf{P_a},\mathbf{P_b},\mathbf{P_{(ab)}}$; the subscripts in equation [7] correspond to the matrices in equation [5]. In the remainder, the ASCA submodels in equation [7] will be indicated as ‘submodel (a)’, ‘submodel (b)’ and ‘submodel (ab)’. Each SCA-model consists of a predefined number of components indicated by $R_\mathbf{a}$, $R_\mathbf{b}$ and $R_\mathbf{(ab)}$ respectively; **E** is a matrix in which the residuals of all submodels of the ASCA-model are collected: $\mathbf{E} = \mathbf{E_a} + \mathbf{E_b} + \mathbf{E_{(ab)}}$, where $\mathbf{E_a}$ are the residuals of submodel (a), etc.

In the ASCA model, all contributions to the variation by the factors and interactions in the experimental design given in Figure 32 are disentangled: the model is exhaustive for this design. However, in its current definition it is not unique: each SCA(-P) submodel can be rotated without a loss of fit of the model. To identify a unique model, each loading matrix is defined to have an orthogonal column space, such that for example $\mathbf{P}_{\mathbf{a}}^{\mathbf{T}}\mathbf{P}_{\mathbf{a}} = \mathbf{I}_{R_{\mathbf{a}}}$ (however the column spaces of the different loading matrices are not mutually orthogonal, such that for example: $\mathbf{P}_{\mathbf{a}}^{\mathbf{T}}\mathbf{P}_{\mathbf{b}} \neq \mathbf{I}_{R_{\mathbf{a}}}$ for $R_{\mathbf{a}} = R_{\mathbf{b}}$). Each subsequent component in each submodel is defined such that it describes as much of the variation in the data as possible that is not yet described by other components. and thereby a unique solution for the model is obtained (that is equal to the Singular Value Decomposition (SVD) solution of each matrix $\mathbf{X}_{\mathbf{a}},\mathbf{X}_{\mathbf{b}},\mathbf{X}_{(\mathbf{ab})}$).

By interpreting the differences between the loadings of the different submodels ($\mathbf{P}_{\mathbf{a}}$, $\mathbf{P}_{\mathbf{b}}$ and $\mathbf{P}_{(\mathbf{ab})}$), the relationship between the variables can be identified for every contribution to the variation. The ASCA model also contains submodel (ab). This submodel describes the variation of the ANOVA ‘error’ term. In ASCA this contribution to the variation is regarded as a source of usable information about the experiment and the collected data.

The column space of the scores of any SCA model lies within the column space of the data the model is fitted on (for details see Appendix 2). This means that by defining matrices $\mathbf{X}_{\mathbf{a}},\mathbf{X}_{\mathbf{b}},\mathbf{X}_{(\mathbf{ab})}$ that have orthogonal column spaces in equation [5], the column spaces of the submodels fitted on these matrices in equation [7] are also mutually orthogonal, which ensures that they can be interpreted independent of each other (15).

Although ASCA is explained here with the two-way ANOVA without replication, any ANOVA equation can be used in the ASCA approach; thus ASCA is a general method for the analysis of multivariate data with an underlying experimental design.

6.2.3 Properties of ASCA

In this section, the properties of ASCA are compared to those of ANOVA and SCA. Since both ANOVA and SCA are the building blocks of ASCA, the properties of ASCA are very similar to those of ANOVA and SCA. This section deals with the ability to merge factors and interactions into one term in ANOVA (which is not directly generalizable in ASCA) and with constraining the ASCA model such that the SCA model is obtained. Furthermore, several criteria for the quality of an ASCA model are defined.

6.2.3.1 ANOVA and ASCA

The ANOVA given in equation [1] is not the only ANOVA model that can be chosen for the experimental design in Figure 32. However, it is the most elaborate: all contributions to the variation that can be expected from the experimental design are separately described by a parameter in equation [1]. In specific cases not all separate factors in the design are of interest. Then multiple factors and interactions in the ANOVA equation can be collected into one ANOVA parameter.

For example, quite possibly factor β of the design in Figure 32 might not be independently of interest, but only in combination with $(\alpha\beta)_{cd}$. The resulting ANOVA equation can then be obtained from equations [1] and [4] as described in equation [8] (index j is again omitted for simplicity).

[8] $$x_{cd} = \mu + \alpha_c + \left[\beta_d + (\alpha\beta)_{cd}\right] = \mu + \alpha_c + (\beta + \alpha\beta)_{cd}$$

This equation shows that by collecting several terms together, an ANOVA model is generated in which the variation is described by a parameter α_c that contains the variation of factor α, and a parameter $(\beta + \alpha\beta)_{cd}$ that describes the remainder of the variation in x_{cd}. For multivariate data, equation [8] can also be written in matrices (analogous to equation [5]) such that equation [9] is obtained.

[9] $$\mathbf{X} = \mathbf{1m}^{\mathbf{T}} + \mathbf{X_a} + \left(\mathbf{X_b} + \mathbf{X_{(ab)}}\right) = \mathbf{1m}^{\mathbf{T}} + \mathbf{X_a} + \mathbf{X_{b+(ab)}}$$

where $\mathbf{X}_{\mathbf{b}+(\mathbf{ab})}$ is a matrix of dimensions equal to **X** that contains the estimates of the ANOVA parameter $(\beta+\alpha\beta)_{cd}$. These estimates can be determined combining equation [8] with Table 3: $(x_{.d}-x_{..})+(x_{cd}-x_{c.}-x_{.d}+x_{..})=(x_{cd}-x_{c.})$. For notation see this table.

An ASCA model can be made of the ANOVA model given in equation [9] using the method described in this paper. This ASCA model is given by equation [10].

$$\mathbf{X}=\mathbf{1m}^{\mathbf{T}}+\mathbf{T_a P_a^T}+\mathbf{T}_{\mathbf{b}+(\mathbf{ab})}\mathbf{P}_{\mathbf{b}+(\mathbf{ab})}^{\mathbf{T}}+\mathbf{E} \qquad [10]$$

where the residuals **E** from equation [10] are generally different from the residuals **E** from equation [7] (the same symbol is chosen for all residuals because of simplicity of notation).

The (b+(ab)) submodel in equation [10] can in general not be obtained from an addition of the (b) and (ab) submodels in equation [7]. Also usually no rotation can be defined that transforms the (b) and (ab) submodels into the (b+(ab)) submodel. The approximation of the variation in $\mathbf{X}_{\mathbf{b}+(\mathbf{ab})}$ is generally different from the independent approximations of the variation in $\mathbf{X_b}$ and in $\mathbf{X}_{(\mathbf{ab})}$. The difference between both models is that in equation [10] the variations related to factor β and interaction $\alpha\beta$ are confounded, while in equation [7] both contributions to the variation are disentangled and separately described by the two loading matrices. However, the models $\mathbf{T_a P_a^T}$ in equations [7] and [10] are equal.

6.2.3.2 SCA and ASCA

The relationship between SCA and ASCA is described in this section. The score matrix **T** of the SCA model in equation [3] can be decomposed taking into account the experimental design (analogous to **X** in equation [5]) into multiple matrices: $\mathbf{T}=\mathbf{T_a}+\mathbf{T_b}+\mathbf{T}_{(\mathbf{ab})}$.

Table 4 Construction of matrices $\mathbf{T_a}$, $\mathbf{T_b}$ and $\mathbf{T_{(ab)}}$ in equation [11] from T in equation [3]. Each element in T can be represented as t_{cdr}; $t_{c.r}$ is the average score of level *c* for all levels *d* and component *r*; $t_{.dr}$ is the average score of level *d* for all levels *c* and component *r*, $t_{..r}$ is the average score of all levels *d* and all levels *c* for component *r*.

Matrix	Construction
$\mathbf{T_a}$	$t_{c.r} - t_{..r}$
$\mathbf{T_b}$	$t_{.dr} - t_{..r}$
$\mathbf{T_{(ab)}}$	$t_{cdr} - t_{c.r} - t_{.dr} + t_{..r}$

For this decomposition the estimates in Table 4 can be used, which impose certain constraints on matrices $\mathbf{T_a}$, $\mathbf{T_b}$ and $\mathbf{T_{(ab)}}$. This is described in equation [11] (the data in **X** is assumed to be mean-centered).

$$\mathbf{X} = \left(\mathbf{T_a} + \mathbf{T_b} + \mathbf{T_{(ab)}}\right)\mathbf{P^T} + \mathbf{E} = \mathbf{T_a}\mathbf{P^T} + \mathbf{T_b}\mathbf{P^T} + \mathbf{T_{(ab)}}\mathbf{P^T} + \mathbf{E} \quad [11]$$

The approach in equation [11] is the PC-ANOVA approach that is described in the introduction (5, 13). Although the scores in the different matrices reflect the experimental design, all factors and interactions in the design are expressed on the same subspace. This means that in PC-ANOVA, the relative importance of a measured variable cannot be attributed to any of the factors (or interactions) in the experimental design: all contributions to the total variation in the data are confounded within one loading matrix $\mathbf{P}$. This hampers the amount of information that can be obtained from the data.

Comparison of equation [11] to equation [7] shows that PC-ANOVA and ASCA differ in that the ASCA fits a separate set of loadings for each factor and interaction in the experimental design (although for simplicity the same symbols are used for the scores of both models). Comparison of these loadings will indicate which measured variables are important in which factor or interaction in the design. ASCA and PC-ANOVA (and therefore SCA) are equal to each other when only one factor or interaction contributes to the variation in the data: e.g. when only experimental factor α causes variation in the data, score matrices $\mathbf{T_b}$ and $\mathbf{T}_{(\mathbf{ab})}$ contain only zeros and the ASCA model consists only of submodel (a), such that $\mathbf{P_a}$ is equal to $\mathbf{P}$. Also, ASCA and PC-ANOVA are equal to each other when the variation of each factor and interaction occurs in the same subspace, such that $\mathbf{P_a} = \mathbf{P_b} = \mathbf{P}_{(\mathbf{ab})} = \mathbf{P}$. However, both situations are unlikely to occur in practice.

The maximum number of components that can be fitted for an SCA model (as given in equation [3]) is easily determined: the rank of a mean-centered matrix **X** is equal to $\min(N-1, J)$, provided that **X** has full row or column rank (which is generally the case for real data); therefore the maximum number of SCA components that can be fitted for **X** is also $\min(N-1, J)$.

The maximum number of components that can be fitted for each ASCA submodel is limited: the rank of each matrix $\mathbf{X_a}, \mathbf{X_b}, \mathbf{X}_{(\mathbf{ab})}$ and therefore the maximum number of components that can be fitted is given in Appendix 3, together with a general method to determine these ranks for any ASCA model.

6.2.3.3 Variation in the ASCA model

The variation of the data explained by the model is used as a quality of fit in ASCA, which is common in multivariate data analysis. The variation explained by the ASCA model can be calculated in multiple ways. Each amount of variation will indicate the quality of the model using a different view. The percentages of explained variation described in this section are all given in Table 5. The notation used in Table 5 is explained in Figure 33.

Table 5 Different criteria for the quality-of-fit (percentage of explained variation) of the ASCA model. The qualities of fit for each submodel are illustrated using the relevant criterion for submodel (ab); $r_{(\mathbf{ab})}$ is the index for component number in submodel (ab).

Submodel	
Total	$ev_{(\mathbf{ab})} = \left(1 - \frac{\left\Vert \mathbf{X}_{(\mathbf{ab})} - \mathbf{T}_{(\mathbf{ab})}\mathbf{P}^{\mathbf{T}}_{(\mathbf{ab})} \right\Vert^{\mathbf{2}}}{\left\Vert \mathbf{X}_{(\mathbf{ab})} \right\Vert^{\mathbf{2}}}\right) \times 100\%$
Per component	$ev_{(\mathbf{ab}),r_{(\mathbf{ab})}} = \left(1 - \frac{\left\Vert \mathbf{X}_{(\mathbf{ab})} - \mathbf{t}_{(\mathbf{ab}),r_{(\mathbf{ab})}}\mathbf{p}^{\mathbf{T}}_{(\mathbf{ab}),r_{(\mathbf{ab})}} \right\Vert^{\mathbf{2}}}{\left\Vert \mathbf{X}_{(\mathbf{ab})} \right\Vert^{\mathbf{2}}}\right) \times 100\%$
Per subset of samples	$ev_{(\mathbf{ab}),c} = \left(1 - \frac{\left\Vert \mathbf{X}_{(\mathbf{ab}),c} - \mathbf{T}_{(\mathbf{ab}),c}\mathbf{P}^{\mathbf{T}}_{(\mathbf{ab})} \right\Vert^{\mathbf{2}}}{\left\Vert \mathbf{X}_{(\mathbf{ab}),c} \right\Vert^{\mathbf{2}}}\right) \times 100\%$
Per subset of samples per component	$ev_{(\mathbf{ab}),cr_{(\mathbf{ab})}} = \left(1 - \frac{\left\Vert \mathbf{X}_{(\mathbf{ab}),c} - \mathbf{t}_{(\mathbf{ab}),cr_{(\mathbf{ab})}}\mathbf{p}^{\mathbf{T}}_{(\mathbf{ab}),r_{(\mathbf{ab})}} \right\Vert^{\mathbf{2}}}{\left\Vert \mathbf{X}_{(\mathbf{ab}),c} \right\Vert^{\mathbf{2}}}\right) \times 100\%$
Total model	
Total	$ev_{\mathbf{tot}} = \left(1 - \frac{\left\Vert \mathbf{X} - \mathbf{1m}^{\mathbf{T}} - \mathbf{T}_{\mathbf{a}}\mathbf{P}^{\mathbf{T}}_{\mathbf{a}} - \mathbf{T}_{\mathbf{b}}\mathbf{P}^{\mathbf{T}}_{\mathbf{b}} - \mathbf{T}_{(\mathbf{ab})}\mathbf{P}^{\mathbf{T}}_{(\mathbf{ab})} \right\Vert^{\mathbf{2}}}{\left\Vert \mathbf{X} - \mathbf{1m}^{\mathbf{T}} \right\Vert^{\mathbf{2}}}\right) \times 100\%$
Total per subset of samples	$ev_{\mathbf{tot},c} = \left(1 - \frac{\left\Vert \mathbf{X}_c - \mathbf{1m}^{\mathbf{T}} - \mathbf{T}_{\mathbf{a},c}\mathbf{P}^{\mathbf{T}}_{\mathbf{a}} - \mathbf{T}_{\mathbf{b},c}\mathbf{P}^{\mathbf{T}}_{\mathbf{b}} - \mathbf{T}_{(\mathbf{ab}),c}\mathbf{P}^{\mathbf{T}}_{(\mathbf{ab})} \right\Vert^{\mathbf{2}}}{\left\Vert \mathbf{X}_c - \mathbf{1m}^{\mathbf{T}} \right\Vert^{\mathbf{2}}}\right) \times 100\%$

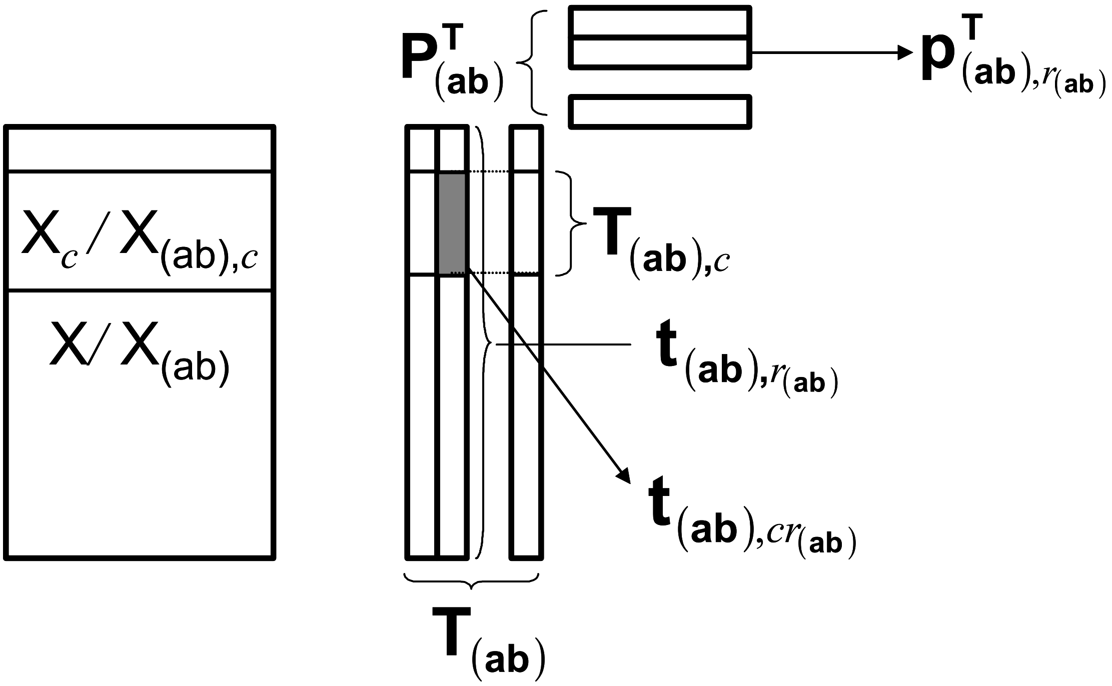
$\mathbf{P}_{(\mathbf{ab})}^{\mathbf{T}}$
$\mathbf{p}_{(\mathbf{ab}),r_{(\mathbf{ab})}}^{\mathbf{T}}$
$X_c / X_{(ab),c}$
$X/X_{(ab)}$
$\mathbf{T}_{(\mathbf{ab}),c}$
$\mathbf{t}_{(\mathbf{ab}),r_{(\mathbf{ab})}}$
$\mathbf{t}_{(\mathbf{ab}),cr_{(\mathbf{ab})}}$
$\mathbf{T}_{(\mathbf{ab})}$

Figure 33 Explanation of the symbols used in Table 5

The quality of fit of submodel (ab) is given by $ev_{(\mathbf{ab})}$ in Table 5: this value is generally calculated to determine the quality of fit of an SCA model. Since the column space of each loading matrix in the ASCA model is orthogonal, $ev_{(\mathbf{ab}),r_{(\mathbf{ab})}}$ can also be calculated to determine the explained variation *per* submodel component (such that $\sum_{r_{(\mathbf{ab})}=1}^{R_{(\mathbf{ab})}} ev_{(\mathbf{ab}),r_{(\mathbf{ab})}} = ev_{(\mathbf{ab})}$ for submodel (ab)). The quality of fit $ev_{(\mathbf{ab}),c}$ of submodel (ab) to a submatrix $\mathbf{X}_{(\mathbf{ab}),c}$ of matrix $\mathbf{X}_{(\mathbf{ab})}$ (e.g. one that contains only a subset of the measurements, that are related to level c of factor α) can also be calculated. This value $ev_{(\mathbf{ab}),c}$ can also be calculated *per* component, such that $\sum_{r_{(\mathbf{ab})}=1}^{R_{(\mathbf{ab})}} ev_{(\mathbf{ab}),cr_{(\mathbf{ab})}} = ev_{(\mathbf{ab}),c}$. Note that these values can also be calculated for submodel (a) and submodel (b) in the ASCA model. The fit of the total ASCA model (for all submodel simultaneously) is given by $ev_{\mathbf{tot}}$. This total fit can also be determined for a submatrix $\mathbf{X}_c$ of matrix **X**, such that $ev_{\mathbf{tot},c}$ is obtained.

6.3 Case studies

To give an impression of the versatility of the use of ASCA for practical applications, three case studies are given in this section. All studies are on data obtained from time-resolved metabolomics studies on mammals. In these studies body fluids of multiple animals are collected at different points in time. The chemical composition of these body fluids is monitored using Nuclear Magnetic Resonance (NMR) spectroscopy.

The description of these case studies is largely limited to the selection of an ASCA model using the design of the experiment. The results of analyses of these datasets with ASCA have been reported elsewhere (1, 7, 16, 19).

6.3.1 Time-resolved metabolomics data

Time-resolved metabolomics experiments are generally performed to monitor the variation in response of the metabolism to an exogenous perturbation. The two factors related to this are the factor 'time' that is given the symbol α and index $1 \ldots k \ldots K$ and the factor 'treatment' that has symbol β and index $1 \ldots h \ldots H$.

In these experiments, generally the factors α and β are crossed, such that each treatment is measured at each time-point. To be able to distinguish the interaction between α and β from the contribution of individual animals, the experiment is generally repeated on multiple animals for each combination of h and k. These animals are indicated by factor γ and index $(1 \ldots i \ldots I)$.The multiple variables in the data are the different NMR channels in the data $(1 \ldots j \ldots J)$.

Before, α and β were used to indicate experimental factors that were not specified. In the current and the next section, α and β will be used exclusively for 'time' and 'treatment group' respectively.

6.3.2 Normal variation

The first case study deals with an experiment where no external influences are imposed on the metabolism. Urine is collected from I different monkeys at K points in time. The monkeys are not divided into treatment groups, therefore design factor β is not present in this design. This design is schematically depicted in Figure 34.

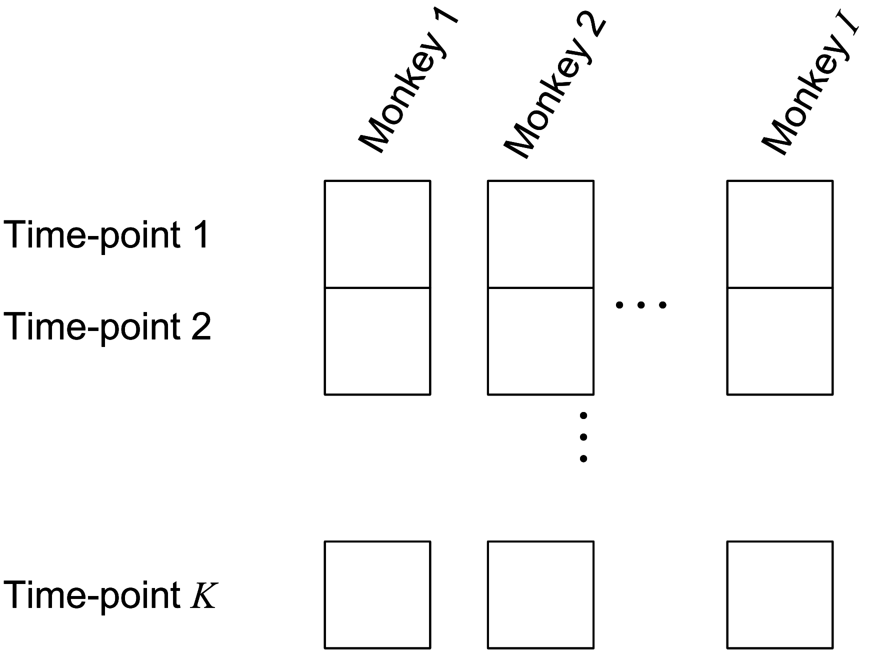

Figure 34 Design of the normality experiment. Each square corresponds to a sample in the design and therefore to an NMR spectrum.

There is no common starting time in the design, since the data values of different animals are unrelated at each time point (the only variation in the data is 'normal' biological variation that is different for each monkey). The factor 'time' (α) can not be usefully analyzed in this particular design and therefore the total variation in the data can be separated into variation between the monkeys that is constant in time and variation in time that is unique for each monkey. The one-way ANOVA in equation [12] is selected for the analysis of this dataset.

[12] $$x_{ikj} = \mu_j + \gamma_{ij} + (\alpha\gamma)_{ikj}$$

The ASCA model obtained from equation [12] is given in equation [13].

[13] $$\mathbf{X} = \mathbf{1m}^{\mathbf{T}} + \mathbf{T_g P_g^T} + \mathbf{T_{(ag)} P_{(ag)}^T} + \mathbf{E}$$

where submodel (g) describes the variation of factor ‘individual animal’ and submodel (ag) describes the dynamic variation of all monkeys. The scores $\mathbf{T_g}$ are highly structured and vary only between the animals and not between time-points.

The analysis of this dataset using the ASCA model in equation [13] (which is also known as *Multilevel* Simultaneous Component Analysis, or MSCA) has been described previously (7). It has been demonstrated that disentangling the variation between the monkeys from the dynamic variation of each individual monkey greatly increases the amount of information that can be extracted.

6.3.3 Bromobenzene

The next case deals with a toxic insult of the metabolism by bromobenzene. In this study groups of rats are treated with different doses of this toxic compound. At different measurement time-points, three rats are randomly selected from each treatment group. Urine is collected from these rats and they are subsequently removed from the experiment. This design is illustrated in Figure 35. The figure shows that the animals are nested within factors time-point and treatment group; hence they are indicated by index i_{hk}. To analyze this experimental design, a two-way ANOVA with replication is chosen. This model is given in equation [14].

[14] $$x_{hi_{hk}kj} = \mu_j + \alpha_{kj} + \beta_{hj} + (\alpha\beta)_{hkj} + (\alpha\beta\gamma)_{hi_{hk}kj}$$

Because each individual animal is measured only once during the BB study, the individual related variation cannot be divided into a part that is constant in time and a dynamic part: both are expressed by one parameter $(\alpha\beta\gamma)_{hi_{hk}kj}$. The ASCA model corresponding to ANOVA equation [14] is given in equation [15].

$$\mathbf{X} = \mathbf{1m}^{\mathbf{T}} + \mathbf{T_a P_a^T} + \mathbf{T_b P_b^T} + \mathbf{T_{(ab)} P_{(ab)}^T} + \mathbf{T_{(abg)} P_{(abg)}^T} + \mathbf{E} \qquad [15]$$

where submodel (a) describes the variation of factor ‘time’ , submodel (b) describes the variation of the factor ‘treatment group’, submodel (ab) describes the interaction between ‘time’ and ‘treatment group’ and submodel (abg) describes the biological variation in the data.

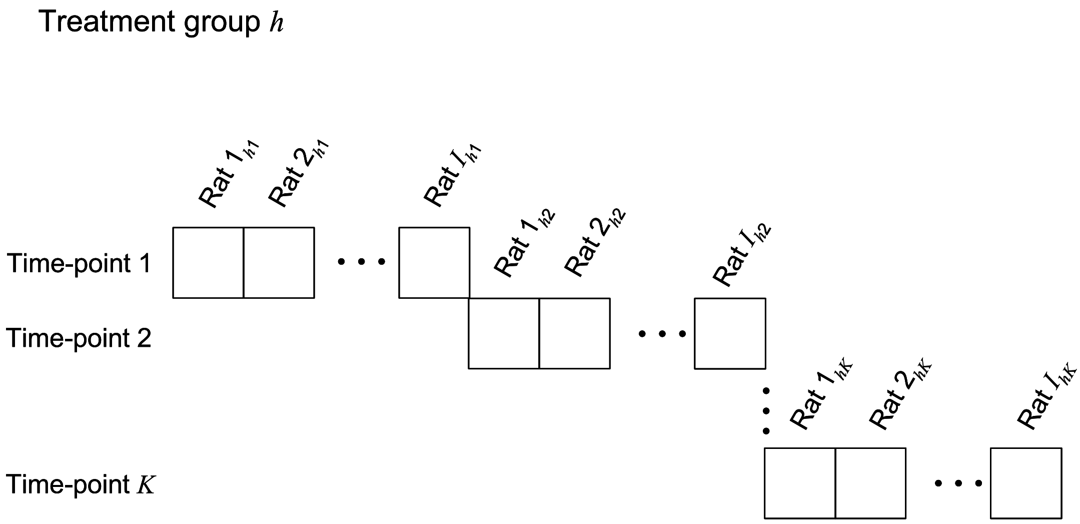

Figure 35 Design of the bromobenzene experiment for one treatment group. Each square corresponds to a sample in the design and therefore to an NMR spectrum. This figure represents the design of one treatment group *h*.

In the ASCA model given in equation [15] the variations related to all factors and interactions that are present in the data are disentangled: this ASCA model is most comprehensive for this design, because each contribution to the variation is expressed in a different subspace spanned by the different loading matrices.

Alternatively, the contributions by the two ANOVA parameters corresponding to the treatment groups (β_{hj} and $(\alpha\beta)_{hkj}$) can be expressed in one submodel, such that these contributions to the variation are confounded. In this specific case, the advantage of confounding is that the treatment group-specific variation can be analyzed by interpreting only one figure. When the decision for this combination is made, the ASCA model in equation [16] is obtained.

[16] $$\mathbf{X} = \mathbf{1m}^{\mathbf{T}} + \mathbf{T_a P_a^T} + \mathbf{T}_{(\mathbf{b+ab})}\mathbf{P}^{\mathbf{T}}_{(\mathbf{b+ab})} + \mathbf{T}_{(\mathbf{abg})}\mathbf{P}^{\mathbf{T}}_{(\mathbf{abg})} + \mathbf{E}$$

where submodel (b+ab) describes the variation of the collected parameters $(\beta + \alpha\beta)_{hkj}$, which in this experiment contains all variation related to the treatment groups.

The dataset contains four treatment groups: one group is administered no toxin, the other three groups are given a low, medium and high dose of bromobenzene respectively. The scores of submodel (b+ab) in equation [16] are given in Figure 36. These scores are focused on describing the different reactions in time of the treatment groups. The scores indicate a quantitative effect of bromobenzene. The scores increase for an increasing dose of toxin. The scores also show that the metabolism of the low and medium dosed rats return to the range of the scores of the rats that are administered no Bromobenzene (the low dosed rats after 24 hours and the medium dosed rats after 48 hours). The metabolism of the high dosed rats however, remains influenced by Bromobenzene throughout the experiment: the high dose of Bromobenzene is higher than the homeostatic capacity of the rats. More results of the ASCA analysis on the Bromobenzene dataset have been described elsewhere (19).

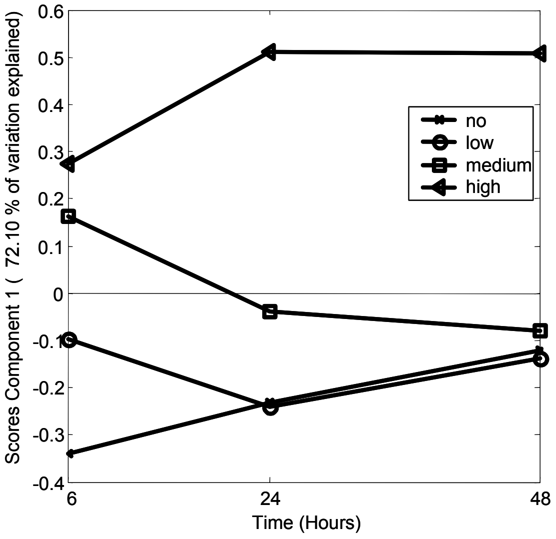

Figure 36 submodel (b+(ab)) scores for the bromobenzene data

6.3.4 Osteoarthritis and Vitamin C

The third case study deals with disease development. A control group is compared to a selected strain of guinea pigs that develops osteoarthritis (OA). The latter group of guinea pigs is subdivided into different treatment groups to which different doses of vitamin C are administered. Urine is collected from the same guinea pigs at multiple time-points. Vitamin C is expected to have an effect on the development of OA in these guinea pigs.

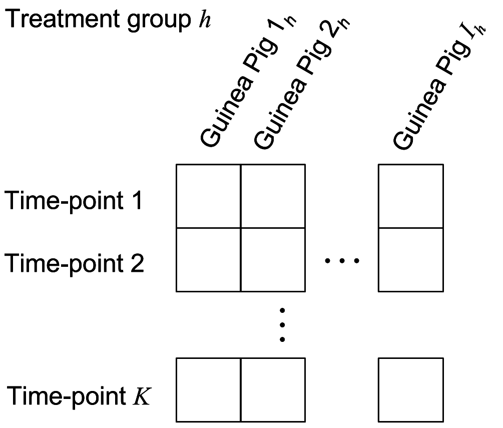

Figure 37 Design of the Osteoarthritis experiment for one treatment group. Each square corresponds to a sample in the design and therefore to an NMR spectrum. This figure represents the design of one treatment group *h*.

This experimental design is schematically depicted in Figure 37 and the exhaustive ANOVA model for this design is given in equation [17]. In this equation the index for the individual animal i only contains a subscript h, because in this design the animals are crossed with the measurement time-points, which is also shown in the figure.

$$x_{hi_{hk}kq} = \mu_j + \alpha_{kj} + \beta_{hj} + (\alpha\beta)_{hkj} + (\beta\gamma)_{hi_h j} + (\alpha\beta\gamma)_{hi_h kj} \qquad [17]$$

where $(\beta\gamma)_{hi_h j}$ is parameter for the interaction of the factor 'individual animal' with 'treatment group' that is constrained as $\sum_{i_h=1}^{I_h} (\beta\gamma)_{hi_h j} = 0 \;\forall\; h, j$ and $(\alpha\beta\gamma)_{hi_h kj}$ is the parameter for the contribution of the biological variation that is constrained both as $\sum_{i_h=1}^{I_h} (\alpha\beta\gamma)_{hi_h kj} = 0 \;\forall\; h, k, j$ and $\sum_{k=1}^{K} (\alpha\beta\gamma)_{hi_h kj} = 0 \;\forall\; i_h, j$: this split-up of the individual-specific variation into two contributions cannot be performed in the Bromobenzene experiment, since each rat is removed from that experiment after its urine is collected.

When ANOVA is used for inference, $(\beta\gamma)_{hi_h j}$ and $(\alpha\beta\gamma)_{hi_h kj}$ are usually added into one 'error' term (analogous to equation [8]). In ASCA this leads to the confounding of the variation related to both terms and therefore, unless it is explicitly decided to do otherwise, they are separated from each other.

From equation [17] an ASCA model can be made using the procedure described before. This model, in which all contributions to the variation of $x_{hi_{hk}kj}$ are disentangled, is given in equation [18].

$$\mathbf{X} = \mathbf{1m}^\mathbf{T} + \mathbf{T_a P_a^T} + \mathbf{T_b P_b^T} + \mathbf{T_{(ab)} P_{(ab)}^T} + \mathbf{T_{(bg)} P_{(bg)}^T} + \mathbf{T_{(abg)} P_{(abg)}^T} + \mathbf{E} \qquad [18]$$

Also, analogous to the bromobenzene experiment the ASCA model in equation [16] can be selected (again, despite the fact that different contributions to the variation are confounded within the submodels). The results of an analysis of the Osteoarthritis dataset with the model in equation [16] can be found in (1).

6.4 Relationship of ASCA to other methods

Several methods to use the experimental design in multivariate data analysis have been proposed in the literature. In the introduction MANOVA and APLSR were already mentioned, but these are not further discussed here.

In this section two of these methods, SMART analysis (20) and Principal Response Curves (21), are discussed together with their similarities and differences to ASCA.

Both methods employ a specific problem-based parameterization of the available data, to focus its analysis on the variation that is of interest to the experimenter and discard unwanted variation from the fitted component model. The notation used in this section is equal to that used in the ‘case studies’ section. Both methods are tailor-made for the analysis of data that has the same structure as a time-resolved metabolomics dataset (e.g. the Osteoarthritis data that was described before).

6.4.1 SMART Analysis

SMART is a method that has been developed specifically for the analysis of time-resolved metabolomics data. There is often biological variation present between the animals prior to the start of the experiment. The aspect of SMART that is discussed here is that each treatment-specific trajectory is expressed as a deviation from the average pre-dose spectrum of each treatment. The parametrization chosen in SMART analysis is given in equation [19].

[19] $$x_{hi_hkj} = \beta_{hj} + (\alpha\beta)_{hkj} + (\alpha\beta\gamma)_{hi_hkj}$$

In SMART, specific constraints are chosen for the parameters in equation [19] that are different from the ‘usual constraints’: β_{hj} is constrained as the average pre-dose value (k=1) of variable j for treatment group h; parameter $(\alpha\beta)_{hkj}$ is constrained as $(\alpha\beta)_{h1j} = 0 \, \forall \, h, j$ and parameter $(\alpha\beta\gamma)_{hi_hkj}$ is constrained as $\sum_{i_h=1}^{I_h} (\alpha\beta\gamma)_{hi_hkj} = 0 \, \forall \, h, k, j$. The estimates of the terms in equation [19] (in combination with the constraints defined on the parameters) are given in equation [20].

[20] $$x_{hi_hkj} = x_{h.1j} + (x_{h.kj} - x_{h.1j}) + (x_{hi_hkj} - x_{h.kj})$$

where $x_{h.1j}$ is the average value of variable j for treatment group h at the beginning of the experiment (k=1); $x_{h.kj}$ is the value of variable j for treatment group h at time-point k, averaged over all animals within the treatment group.

Equation [20] shows that in SMART the estimates of $(\alpha\beta)_{hkj}$ are zero at k=1 and that β_{hj} contains the average pre-dose spectrum for each group. Matrices that are constructed from the estimates in equation [20] do not have orthogonal column spaces (such that the relation between equations [5] and [6] does not hold for SMART). Keun *et al.* only use the estimates of $(\alpha\beta)_{hkj}$ given in equation [20] to construct a component model.

6.4.2 Principal Response Curves

Another method that uses a problem–specific parameterization to remove undesired variation from the model is Principal Response Curves (PRC) Analysis (21). Although PRC is developed in ecology, the design of the experiments for which the method is tailor-made can be generalized to time-resolved metabolomics experiments.

PRC models the treatment groups as a deviation from the control group, to enable a better view on the treatment-group related variation in the data. The ANOVA-parameterization used in PRC is given in equation [21].

[21] $$x_{hi_hkj} = \alpha_{kj} + (\alpha\beta)_{hkj} + (\alpha\beta\gamma)_{hi_hkj}$$

where α_{kj} is the parameter of factor 'time' for variable j and is constrained in PRC to describe the variation of the control group (h=1); $(\alpha\beta)_{hkj}$ is the parameter for the interaction between treatment group h and time-point k for variable j and is constrained as $(\alpha\beta)_{1kj} = 0 \; \forall \; k, j$; $(\alpha\beta\gamma)_{hi_hkj}$ is the parameter for the error of animal i_h at measurement time-point k for variable j and constrained as $\sum_{i_h=1}^{I_h} (\alpha\beta\gamma)_{hi_hkj} = 0 \; \forall \; h,k,j$.

The terms in equation [21] (in combination with the defined constraints on the parameters) can be estimated using equation [22].

[22] $$x_{hi_hkj} = x_{1.kj} + (x_{h.kj} - x_{1.kj}) + (x_{hi_hkj} - x_{h.kj})$$

where $x_{1.kj}$ is the value of variable j for the control group (h=1) at time-point k; $x_{h.kj}$ is the value of variable j for treatment group h at time-point k, averaged over all animals within the treatment group.

The estimates in equation [22] impose the constraints defined for the parameters, but will not lead to orthogonality of the column spaces of matrices constructed from these estimates (viz. equations [5] and [6]), which means that the variation of each estimate cannot be individually interpreted. Only the estimates of $(\alpha\beta)_{hkj}$ are used in PRC and the variation of parameters α_{kj} and $(\alpha\beta\gamma)_{hi_hkj}$ is discarded from the model. The PRC model consists of an SCA model on the matrix (of dimensions equal to **X**) constructed from the estimates of $(\alpha\beta)_{hkj}$.

6.5 Extensions of the ASCA model

Thus far ASCA has been presented as an extension of the PCA model. However, the idea behind ASCA can be used in many more applications of multivariate data analysis. ASCA should therefore not only be seen as a novel data analysis method as such, but also as a framework for the analysis of multivariate data originating from an experimental design.

First of all, the SCA methodology can be used more exhaustively in ASCA. The SCA-P model used here has properties that are identical to PCA. Aside from SCA-P multiple alternative versions have been proposed (4). The increasingly constrained SCA-PARAFAC2 (SCA-PF2), SCA-IND and SCA-ECP models can replace the SCA-P submodels straightforwardly when desired.

6.5.1 Multiway-ASCA

Some ANOVA parameters vary for levels of multiple factors in the experimental design: e.g. $(\alpha\beta)_{cdj}$ in equation [4] varies both for levels c of factor α and for levels d of factor β. In ASCA the variation of $(\alpha\beta)_{cdj}$ is modeled by the component scores that describe the variation of factors α and β (with levels c and d) and the loadings that describe the original variables j. An elementwise expression of submodel (ab) from equation [7] is given in equation [23].

$$[23] \quad x_{(\mathbf{ab}),cdj} = \sum_{r_{(\mathbf{ab})}=1}^{R_{(\mathbf{ab})}} t_{(\mathbf{ab}),cdr_{\mathbf{ab}}} p_{(\mathbf{ab}),jr_{\mathbf{ab}}} + e_{(\mathbf{ab}),cdj}$$

where $x_{(\mathbf{ab}),cdj}$ is the (cd x j)'th element of matrix $\mathbf{X}_{(\mathbf{ab})}$, the term $\sum_{r_{\mathbf{ab}}=1}^{R_{\mathbf{ab}}} t_{(\mathbf{ab}),cdr_{\mathbf{ab}}} p_{(\mathbf{ab}),jr_{\mathbf{ab}}}$ is the SCA estimate of this element and $e_{(\mathbf{ab}),cdj}$ are the residuals of the model.

In equation [23] the scores $t_{\mathbf{ab},cdr_{\mathbf{ab}}}$ can vary freely for each c and d. As an alternative for SCA, a more constrained three-way model (e.g. PARAFAC (22-24) can be selected for the analysis of $\mathbf{X}_{(\mathbf{ab})}$(26). Then instead of equation [23], equation [24] is obtained.

$$[24] \quad x_{(\mathbf{ab}),cdj} = \sum_{r_{(\mathbf{ab})}=1}^{R_{(\mathbf{ab})}} t_{(\mathbf{ab})\mathbf{1},cr_{\mathbf{ab}}} t_{(\mathbf{ab})\mathbf{2},dr_{\mathbf{ab}}} p_{(\mathbf{ab}),jr_{\mathbf{ab}}} + e_{(\mathbf{ab}),cdj}$$

where the component scores in $t_{(\mathbf{ab})\mathbf{1},cr_{\mathbf{ab}}}$ vary with factor α, the component scores $t_{(\mathbf{ab})\mathbf{2},dr_{\mathbf{ab}}}$ vary with factor β, and $p_{(\mathbf{ab}),jr_{\mathbf{ab}}}$ varies with variable j; all three vary also between components $r_{(\mathrm{ab})}$; $p_{(\mathbf{ab}),jr_{\mathbf{ab}}}$ and $e_{(\mathbf{ab}),cdj}$ are generally different between equations [23] and [24].

Aside from increasing the insight in the mechanisms beneath the variation in the data, this approach can be applied to estimate the interaction between α and β in the absence of replicate measurements. It can be seen as a multivariate extension of the approach proposed by Goodman and Haberman (25).

When appropriate, the generalization of the two-way model in equation [23] to the three-way model in [24] can be extended to any multiway model as well. Which multiway model (e.g. PARAFAC, Tucker3) is used in ASCA is left to the preference of the user (26).

6.5.2 Regression-ASCA

The use of ASCA is not limited to unsupervised data analysis: multivariate regression models can also be used as submodels in ASCA, such that a Regression-ASCA model is obtained. Thereby the different contributions to the variation in **X** can be regressed on another variable **y** (or a set of variables **Y**) such that quantitative external information that is not present in **X** can be used in the data analysis (27). When Principal Component Regression models are selected for each submodel, the model in equation [25] is obtained for the two-way analysis of variance without replication.

[25] $$\mathbf{y} = \mathbf{1}\mathbf{m}^{\mathrm{T}} + \mathbf{T}_{\mathbf{a}}\mathbf{b}_{\mathbf{a}} + \mathbf{T}_{\mathbf{b}}\mathbf{b}_{\mathbf{b}} + \mathbf{T}_{(\mathbf{ab})}\mathbf{b}_{(\mathbf{ab})} + \mathbf{E}$$

Where **y** is the size N vector of regressor variables and $\mathbf{b}_{\mathbf{a}}$, $\mathbf{b}_{\mathbf{b}}$, $\mathbf{b}_{(\mathbf{ab})}$ are the regression coefficients for each submodel in equation [4]; $\mathbf{T}_{\mathbf{a}}$, $\mathbf{T}_{\mathbf{b}}$ and $\mathbf{T}_{(\mathbf{ab})}$ are the ASCA scores as obtained from equation [7].

Using ASCA in a regression context, like in equation [25] can extend the amount of information that can be obtained from a multivariate data analysis. ASCA can be extended (analogous to equation [25]) by any regression method that has component scores on which the ‘usual constraints’ are valid such that the theorem in Appendix 2 applies.

Regression methods can also be used for discriminant analysis (e.g. Partial Least Squares – Discriminant Analysis (28)). In general this is done by regressing the data on design variables. Such a regression can be straightforwardly implemented in a Regression-ASCA model, such that a Discrimination-ASCA is obtained that discriminates the samples in the data based on one or a few experimental design factors.

The preceding shows that any multivariate data analysis model can be used within the ASCA framework. The only prerequisite is that the column spaces of the different submodels (e.g. the scores of the SCA models as described in Appendices 1 and 2) are mutually orthogonal. Also, different types of models can be selected for each ASCA submodel.

6.6 Current limitations of ASCA

ASCA as it is presented in this paper is a data analysis method for parameter estimation. Therefore it should be seen as a more information rich alternative for SCA that can be used when the analyzed multivariate data is obtained from a designed experiment. Currently ASCA cannot be used for inferential purposes and therefore significance testing of the different experimental design factors and interactions is not yet possible within the current methodology: it remains a subject for further research. Also, currently a method for estimating the confidence intervals of the estimated model parameters is lacking. For this resampling methods are under study with which these intervals can be determined.

6.7 Conclusions

By merging SCA and ANOVA into ASCA, a method is obtained that combines the advantages of both methods and removes their disadvantages. The properties of ASCA are easy to derive from the properties of SCA and ANOVA. In limiting cases, ASCA is either equal to SCA or to ANOVA. ASCA is a method that can be straightforwardly implemented by the data analyst and can be used to obtain more information from a multivariate dataset from a designed experiment than with existing methods.

In this paper some case studies are presented. All case studies deal with time-resolved metabolomics experiments. However, ASCA is a general data analysis method that can be applied in many scientific disciplines. The focus of these case studies is on the selection of a proper model and the considerations for merging or separating different design factors and interactions in the ASCA model.

It is shown that PRC and SMART can be seen as methods that use a specific ANOVA parametrization of multivariate data that is different from that used in ASCA. Finally, ASCA can be extended to include more constrained component submodels and multiway submodels and to include multivariate regression or discrimination analysis. Especially this last section shows that although ASCA is presented as a novel data analysis method, it is in fact the basis of a framework for the analysis of multivariate data obtained from any experimental design.

6.8 Appendices

6.8.1 Appendix 1: Orthogonality between the column spaces of the matrices in equation [5]

To prove the orthogonality of their column spaces, the structure of the matrices $\mathbf{X_a}$, $\mathbf{X_b}$ and $\mathbf{X_{(ab)}}$ needs to be determined. Matrix $\mathbf{X}$ is built up from size J row vectors $\mathbf{x}_{cd}^{\mathbf{T}}$ that contain the measurement on level c of factor α and level d of factor β.

The matrix structure is chosen as follows: $\mathbf{X} = \begin{bmatrix} \mathbf{x}_{11}^{\mathbf{T}} \\ \vdots \\ \mathbf{x}_{C1}^{\mathbf{T}} \\ \vdots \\ \mathbf{x}_{1D}^{\mathbf{T}} \\ \vdots \\ \mathbf{x}_{CD}^{\mathbf{T}} \end{bmatrix}$.

Matrices $\mathbf{X}_{\mathbf{a}}$, $\mathbf{X}_{\mathbf{b}}$ and $\mathbf{X}_{(\mathbf{ab})}$ are then structured as follows:

$\mathbf{X}_{\mathbf{a}} = \begin{bmatrix} \mathbf{X}_{\mathbf{a}}^{*} \\ \vdots \\ \mathbf{X}_{\mathbf{a}}^{*} \end{bmatrix}$, where $\mathbf{X}_{\mathbf{a}}^{*} = \begin{bmatrix} \mathbf{x}_{\mathbf{a},1}^{\mathbf{T}} \\ \vdots \\ \mathbf{x}_{\mathbf{a},C}^{\mathbf{T}} \end{bmatrix}$ and $\mathbf{x}_{\mathbf{a},c}^{\mathbf{T}}$ is a size J row vector containing values $(x_{c.1} - x_{..1})$ to $(x_{c.J} - x_{..J})$.

$\mathbf{X}_{\mathbf{b}} = \begin{bmatrix} \mathbf{X}_{\mathbf{b},1}^{*} \\ \vdots \\ \mathbf{X}_{\mathbf{b},D}^{*} \end{bmatrix}$, where $\mathbf{X}_{\mathbf{b},d}^{*} = \begin{bmatrix} \mathbf{x}_{\mathbf{b},d}^{\mathbf{T}} \\ \vdots \\ \mathbf{x}_{\mathbf{b},d}^{\mathbf{T}} \end{bmatrix} = \mathbf{1}\mathbf{x}_{\mathbf{b},d}^{\mathbf{T}}$ where $\mathbf{x}_{\mathbf{b},d}^{\mathbf{T}}$ is a size J row vector containing values $(x_{.d1} - x_{.d1})$ to $(x_{.dJ} - x_{.dJ})$.

$\mathbf{X}_{(\mathbf{ab})} = \begin{bmatrix} \mathbf{X}_{(\mathbf{ab}),1}^{*} \\ \vdots \\ \mathbf{X}_{(\mathbf{ab}),D}^{*} \end{bmatrix}$, where $\mathbf{X}_{(\mathbf{ab}),d}^{*} = \begin{bmatrix} \mathbf{x}_{(\mathbf{ab}),1d}^{\mathbf{T}} \\ \vdots \\ \mathbf{x}_{(\mathbf{ab}),Cd}^{\mathbf{T}} \end{bmatrix}$ and $\mathbf{x}_{(\mathbf{ab}),cd}^{\mathbf{T}}$ is a size J row vector containing values $(x_{cd1} - x_{c.1} - x_{.d1} + x_{..1})$ to $(x_{cdJ} - x_{c.J} - x_{.dJ} + x_{..J})$

From the preceding, together with the constraints in Table 3 it can be concluded that I., II. and III. hold:

I. $\mathbf{X}_{\mathbf{a}}$: $\sum_{c=1}^{C}(x_{c.j} - x_{..j}) = 0 \rightarrow \sum_{d=1}^{D}\sum_{c=1}^{C}(x_{c.j} - x_{..j}) = 0 \rightarrow$

$\mathbf{1}_N^{\mathbf{T}}\mathbf{X}_{\mathbf{a}} = \mathbf{0}_J^{\mathbf{T}}$

II. $\mathbf{X}_{\mathbf{b}}$: $\sum_{d=1}^{D}(x_{.dj} - x_{..j}) = 0 \rightarrow \sum_{c=1}^{C}\sum_{d=1}^{D}(x_{.dj} - x_{..j}) = 0 \rightarrow$

$\mathbf{1}_N^{\mathbf{T}}\mathbf{X}_{\mathbf{b}} = \mathbf{0}_J^{\mathbf{T}}$

III. $\mathbf{X}_{(\mathbf{ab})}$: $\sum_{c=1}^{C}(x_{cdj} - x_{c.j} - x_{.dj} + x_{..j}) = 0$ and $\sum_{d=1}^{D}(x_{cdj} - x_{c.j} - x_{.dj} + x_{..j}) = 0 \rightarrow$

$\sum_{d=1}^{D}\sum_{c=1}^{C}(x_{cdj} - x_{c.j} - x_{.dj} + x_{..j}) = 0 \rightarrow \mathbf{1}_N^{\mathbf{T}}\mathbf{X}_{(\mathbf{ab})} = \mathbf{0}_J^{\mathbf{T}}$

I, II and III ensure that the column spaces of $\mathbf{X}_{\mathbf{a}}$, $\mathbf{X}_{\mathbf{b}}$ and $\mathbf{X}_{(\mathbf{ab})}$ are orthogonal to $\mathbf{1m}^{\mathsf{T}}$. The mutual orthogonality between the column spaces of matrices $\mathbf{X}_{\mathbf{a}}$, $\mathbf{X}_{\mathbf{b}}$ and $\mathbf{X}_{(\mathbf{ab})}$ is demonstrated in the remainder of this appendix.

Submodels (a) and (b).(using I.):

$$\mathbf{X}_{\mathbf{a}}^{\mathsf{T}}\mathbf{X}_{\mathbf{b}} = \begin{bmatrix} \mathbf{X}_{\mathbf{a}}^{*} \\ \vdots \\ \mathbf{X}_{\mathbf{a}}^{*} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \mathbf{X}_{\mathbf{b},1}^{*} \\ \vdots \\ \mathbf{X}_{\mathbf{b},D}^{*} \end{bmatrix} = \sum_{d=1}^{D} \mathbf{X}_{\mathbf{a}}^{*\mathsf{T}} \mathbf{X}_{\mathbf{b},d}^{*} = \sum_{d=1}^{D} \mathbf{X}_{\mathbf{a}}^{*\mathsf{T}} \mathbf{1x}_{\mathbf{b},d}^{\mathsf{T}} = \sum_{d=1}^{D} \left(\mathbf{X}_{\mathbf{a}}^{*\mathsf{T}} \mathbf{1}\right) \mathbf{x}_{\mathbf{b},d}^{\mathsf{T}} = \sum_{d=1}^{D} \left(\mathbf{1}^{\mathsf{T}} \mathbf{X}_{\mathbf{a}}^{*}\right)^{\mathsf{T}} \mathbf{x}_{\mathbf{b},d}^{\mathsf{T}} = \mathbf{0}_{(J\times J)}$$

Submodels (a) and (ab).(using III.):

$$\mathbf{X}_{\mathbf{a}}^{\mathsf{T}}\mathbf{X}_{(\mathbf{ab})} = \begin{bmatrix} \mathbf{X}_{\mathbf{a}}^{*} \\ \vdots \\ \mathbf{X}_{\mathbf{a}}^{*} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \mathbf{X}_{(\mathbf{ab}),1}^{*} \\ \vdots \\ \mathbf{X}_{(\mathbf{ab}),D}^{*} \end{bmatrix} = \sum_{d=1}^{D} \mathbf{X}_{\mathbf{a}}^{*\mathsf{T}} \mathbf{X}_{(\mathbf{ab}),d}^{*} = \sum_{c=1}^{C} \left(\mathbf{1x}_{\mathbf{a},c}^{\mathsf{T}}\right)^{\mathsf{T}} \begin{bmatrix} \mathbf{x}_{(\mathbf{ab}),c1}^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_{(\mathbf{ab}),cD}^{\mathsf{T}} \end{bmatrix} = \sum_{c=1}^{C} \mathbf{x}_{\mathbf{a},c} \mathbf{1}^{\mathsf{T}} \begin{bmatrix} \mathbf{x}_{(\mathbf{ab}),c1}^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_{(\mathbf{ab}),cD}^{\mathsf{T}} \end{bmatrix} = \sum_{c=1}^{C} \mathbf{x}_{\mathbf{a},c} \mathbf{0}_{J}^{\mathsf{T}} = \mathbf{0}_{(J\times J)}$$

Submodels (b) and (ab).(using III.):

$$\mathbf{X}_{\mathbf{b}}^{\mathsf{T}}\mathbf{X}_{(\mathbf{ab})} = \begin{bmatrix} \mathbf{X}_{\mathbf{b},1}^{*} \\ \vdots \\ \mathbf{X}_{\mathbf{b},D}^{*} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \mathbf{X}_{(\mathbf{ab}),1}^{*} \\ \vdots \\ \mathbf{X}_{(\mathbf{ab}),D}^{*} \end{bmatrix} = \sum_{d=1}^{D} \mathbf{X}_{\mathbf{b},d}^{*\ \mathsf{T}} \mathbf{X}_{(\mathbf{ab}),d}^{*} = \sum_{d=1}^{D} \left(\mathbf{1x}_{\mathbf{b},d}^{\mathsf{T}}\right)^{\mathsf{T}} \mathbf{X}_{(\mathbf{ab}),d}^{*} = \sum_{d=1}^{D} \mathbf{x}_{\mathbf{b},d} \mathbf{1}^{\mathsf{T}} \mathbf{X}_{(\mathbf{ab}),d}^{*} = \sum_{d=1}^{D} \mathbf{x}_{\mathbf{b},d} \mathbf{0}_{J}^{\mathsf{T}} = \mathbf{0}_{(J\times J)}$$

The proof in this appendix ensures that equation [6] holds for this model, because:

$$\begin{aligned}
&\|\mathbf{X}\|^2 = \left\|\mathbf{1m}^{\mathsf{T}} + \mathbf{X}_{\mathbf{a}} + \mathbf{X}_{\mathbf{b}} + \mathbf{X}_{(\mathbf{ab})}\right\|^2 = tr\left(\left(\mathbf{1m}^{\mathsf{T}} + \mathbf{X}_{\mathbf{a}} + \mathbf{X}_{\mathbf{b}} + \mathbf{X}_{(\mathbf{ab})}\right)^{\mathsf{T}}\left(\mathbf{1m}^{\mathsf{T}} + \mathbf{X}_{\mathbf{a}} + \mathbf{X}_{\mathbf{b}} + \mathbf{X}_{(\mathbf{ab})}\right)\right) = \\
&= tr\left(\left(\mathbf{1m}^{\mathsf{T}}\right)^{\mathsf{T}}\mathbf{1m}^{\mathsf{T}} + \mathbf{X}_{\mathbf{a}}^{\mathsf{T}}\mathbf{X}_{\mathbf{a}} + \mathbf{X}_{\mathbf{b}}^{\mathsf{T}}\mathbf{X}_{\mathbf{b}} + \mathbf{X}_{(\mathbf{ab})}^{\mathsf{T}}\mathbf{X}_{(\mathbf{ab})}\right) = tr\left(\left(\mathbf{1m}^{\mathsf{T}}\right)^{\mathsf{T}}\mathbf{1m}^{\mathsf{T}}\right) + tr\left(\mathbf{X}_{\mathbf{a}}^{\mathsf{T}}\mathbf{X}_{\mathbf{a}}\right) + tr\left(\mathbf{X}_{\mathbf{b}}^{\mathsf{T}}\mathbf{X}_{\mathbf{b}}\right) + \\
&+ tr\left(\mathbf{X}_{(\mathbf{ab})}^{\mathsf{T}}\mathbf{X}_{(\mathbf{ab})}\right) = \\
&= \left\|\mathbf{1m}^{\mathsf{T}}\right\|^2 + \left\|\mathbf{X}_{\mathbf{a}}\right\|^2 + \left\|\mathbf{X}_{\mathbf{b}}\right\|^2 + \left\|\mathbf{X}_{(\mathbf{ab})}\right\|^2
\end{aligned}$$

6.8.2 Appendix 2: Proof that the least-squares minimization of a constrained model is equal to the least-squares minimization of an unconstrained model of projected data on the space where the constraint is valid.

In this Appendix, proof is given that the results of a constrained SCA-P model can also be obtained by constraining the data on which the analysis is performed. **J** is a projection matrix (**JJ** = **J** and $\mathbf{J^T}$ = **J**), **I** is the identity matrix.

The minimization of a linearly constrained PCA model is given by:

$(\mathbf{T,P}) = \underset{\mathbf{T}\in\Re(\mathbf{J}),\mathbf{P}}{\arg\min}\left\|\mathbf{X - TP^T}\right\|^2$ where $\mathbf{P^TP} = \mathbf{I}$, $\mathbf{T^TT}$ is a diagonal matrix in which the elements are sorted from large to small and $\mathbf{T} = \mathbf{JT}$.

The minimization of a PCA model on constrained data is given by:

$(\tilde{\mathbf{T}},\tilde{\mathbf{P}}) = \underset{\tilde{\mathbf{T}},\tilde{\mathbf{P}}}{\arg\min}\left\|\mathbf{JX} - \tilde{\mathbf{T}}\tilde{\mathbf{P}}^\mathbf{T}\right\|^2$ where $\tilde{\mathbf{P}}^\mathbf{T}\tilde{\mathbf{P}} = \mathbf{I}$ and $\tilde{\mathbf{T}}^\mathbf{T}\tilde{\mathbf{T}}$ is a diagonal matrix in which the elements are sorted from large to small.

The question now is: $(\tilde{\mathbf{T}},\tilde{\mathbf{P}}) \overset{?}{=} (\mathbf{T,P})$

$\left\|\mathbf{X - TP^T}\right\|^2$ can be rewritten as:

$$\left\|\mathbf{X - TP^T}\right\|^2 = \left\|(\mathbf{I - J})\mathbf{X}\right\|^2 + \left\|\mathbf{JX - TP^T}\right\|^2$$

since:

$$\left\|\mathbf{X - TP^T}\right\|^2 = \left\|(\mathbf{I - J})\mathbf{X} + \mathbf{JX - TP^T}\right\|^2 = tr\left(\left((\mathbf{I - J})\mathbf{X} + \mathbf{JX - TP^T}\right)^\mathbf{T}\left((\mathbf{I - J})\mathbf{X} + \mathbf{JX - TP^T}\right)\right) =$$

$$= tr\left(\mathbf{X^T}(\mathbf{I - J})^\mathbf{T}(\mathbf{I - J})\mathbf{X}\right) +$$
$$+ tr\left(\mathbf{X^T}(\mathbf{I - J})^\mathbf{T}\mathbf{JX} + \mathbf{X^T}(\mathbf{I - J})^\mathbf{T}\mathbf{TP^T} + \mathbf{X^T J^T}(\mathbf{I - J})\mathbf{X} + \mathbf{X^T J^T JX} - \mathbf{X^T J^T TP^T} - \mathbf{PT^T JX} + \mathbf{PT^T TP^T}\right)$$

where $\mathbf{X^T}(\mathbf{I - J})^\mathbf{T}\mathbf{TP^T} = \mathbf{X^T}(\mathbf{I - J})^\mathbf{T}\mathbf{JTP^T} = \mathbf{X^T 0TP^T} = \mathbf{0}$ since $\mathbf{T} = \mathbf{JT}$.

hence:

$$\left\|\mathbf{X - TP^T}\right\|^2 = tr\left(\mathbf{X^T}(\mathbf{I - J})^\mathbf{T}(\mathbf{I - J})\mathbf{X}\right) + tr\left(\mathbf{X^T J^T JX} - \mathbf{X^T J^T TP^T} - \mathbf{PT^T JX} + \mathbf{PT^T TP^T}\right) =$$

$$= \left\|(\mathbf{I - J})\mathbf{X}\right\|^\mathbf{2} + \left\|\mathbf{JX - TP^T}\right\|^\mathbf{2}$$

This means that:

$$(\mathbf{T},\mathbf{P}) = \underset{\mathbf{T}\in\Re(\mathbf{J}),\,\mathbf{P}}{\arg\min}\left\|\mathbf{X} - \mathbf{TP}^{\mathbf{T}}\right\|^2 = \underset{\mathbf{T}\in\Re(\mathbf{J}),\,\mathbf{P}}{\arg\min}\left(\left\|(\mathbf{I} - \mathbf{J})\mathbf{X}\right\|^2 + \left\|\mathbf{JX} - \mathbf{TP}^{\mathbf{T}}\right\|^2\right) = \left\|(\mathbf{I} - \mathbf{J})\mathbf{X}\right\|^2 + \underset{\mathbf{T},\mathbf{P}}{\arg\min}\left\|\mathbf{JX} - \mathbf{TP}^{\mathbf{T}}\right\|^2 =$$

$$= \underset{\mathbf{T},\mathbf{P}}{\arg\min}\left\|\mathbf{JX} - \mathbf{TP}^{\mathbf{T}}\right\|^2 = \underset{\tilde{\mathbf{T}},\tilde{\mathbf{P}}}{\arg\min}\left\|\mathbf{JX} - \tilde{\mathbf{T}}\tilde{\mathbf{P}}^{\mathbf{T}}\right\|^2 = \left(\tilde{\mathbf{T}},\tilde{\mathbf{P}}\right)$$

Note that the constraint that **T** should be in the range of **J** becomes redundant when it is estimated from **JX**.

Since all submodels in the ASCA models are based on centering in certain directions and every such centering is a projection step, every submodel can be estimated by an unconstrained SCA-P model on properly centered data. A similar proof of this equality is used in CANDELINC (29).

6.8.3 Appendix 3: Maximal number of components for each submodel

The maximum number of components that can be fitted for each ASCA submodel can be determined from the rank of each of the matrices on the right hand side of equation [5]. For each submodel, J is assumed to be larger than the number of rows of each matrix $\mathbf{1m}^{\mathbf{T}}$, $\mathbf{X}_{\mathbf{a}}$, $\mathbf{X}_{\mathbf{b}}$ and $\mathbf{X}_{(\mathbf{ab})}$.

$\mathbf{1m}^{\mathbf{T}}$: μ_q

All rows of $\mathbf{1m}^{\mathbf{T}}$ are equal, therefore $\Re\left(\mathbf{1m}^{\mathbf{T}}\right) = 1$

Submodel (a): $\left(x_{c.j} - x_{...j}\right)$

Matrix $\mathbf{X}_{\mathbf{a}}$ contains C distinct rows. Due to the centering with respect to the overall mean the row rank of $\mathbf{X}_{\mathbf{a}}$ is decreased with 1.

This means that $\Re(\mathbf{X}_{\mathbf{a}}) = \max(R_{\mathbf{a}}) \overset{C-1>J}{=} C-1$

Submodel (b): $\left(x_{.dj} - x_{...j}\right)$

$\mathbf{X}_{\mathbf{b}}$ contains D distinct rows of length J. Due to the centering with respect to the overall mean the row rank of $\mathbf{X}_{\mathbf{b}}$ is decreased with 1.

This means that $\Re(\mathbf{X}_{\mathbf{b}}) = \max(R_{\mathbf{b}}) \overset{D-1<J}{=} D-1$

Submodel (ab): $\left(x_{cdj} - x_{c.j} - x_{.dj} + x_{..j}\right)$

The rank of $\mathbf{X}_{(\mathbf{ab})}$ can be established by determining the number of linearly independent rows in the matrix. This can be done from the estimates contained in $\mathbf{X}_{(\mathbf{ab})}$: $\left(x_{cdj} - x_{c.j} - x_{.dj} + x_{..j}\right)$. The number of independent rows of matrix $\mathbf{X}_{(\mathbf{ab})}$ is then: $(CD - C - D + 1)$.

This means that $\Re\left(\mathbf{X}_{(\mathbf{ab})}\right) = \max\left(R_{(\mathbf{ab})}\right) \overset{CD-C-D+1<J}{=} CD - C - D + 1$.

When J is larger than any of the ranks of $\mathbf{1m}^{\mathbf{T}}$, $\mathbf{X}_{\mathbf{a}}$, $\mathbf{X}_{\mathbf{b}}$ and $\mathbf{X}_{(\mathbf{ab})}$, the maximum row rank aggregated for all submodels is equal to the row rank of the original data **X**, which is CD (assuming that **X** is of full row rank):

$$1 + (C-1) + (D-1) + (CD - C - D + 1) = CD$$

Elimination of all terms on the left hand side of this equation shows that this equality indeed holds. This means that the total aggregated maximum rank of all ASCA submodels (and therefore the maximum total number of components that can be fitted) is equal to the maximum rank of an SCA model and therefore to the rank of the data matrix **X** for a large J.

6.9 Symbol List

Vectors and Matrices	
X	Data
T	Scores
P	Loading
E	Residuals
$\mathbf{x}^{\mathbf{T}}$	Row vector containing a multivariate measurement
$\mathbf{X}^{*}$	Partition of matrix **X** (used in Appendix 1)
1	Vector of ones
m	Vector containing the means of each j
t,p	Score, loading vector
b	Regression Coefficients for an ASCA regression model

J	Projection Matrix
Indices	
$1\ldots c\ldots C$	Index for the levels of factor α
$1\ldots d\ldots D$	Index for the levels of factor β
$1\ldots r\ldots R$	Index for the components
$1\ldots n\ldots N$	Index for the samples
$1\ldots q\ldots Q$	Index for the sets of samples
$1\ldots j\ldots J$	Index for the variables
$1\ldots i\ldots I$	Index for individual animals
$1\ldots h\ldots H$	Index for treatment groups
$1\ldots k\ldots K$	Index for measurement time-points
Scalars	
x	Data value
$\mu, \alpha, \beta, (\alpha\beta),\ldots$	ANOVA parameters
t , p , e	Score, loading and residual values
ev	Explained variation
Labels	
α,β,γ…	Experimental factors
a,b,(ab),(abg),…	Submodel labels
1,2	Subscripts for the PARAFAC model scores

6.10 References

(1) Smilde, A.K., Jansen, J.J., Hoefsloot, H.C.J., van der Greef, J. and Timmerman, M.E., ANOVA-Simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*, 2005; In Press

(2) Searle, S.R., *Linear Models*. John Wiley & Sons, Inc., New York, 1971

(3) Sokal, R.R. and Rohlf, F.J., *Biometry*. W.H. Freeman and company, San Francisco, 1995

(4) Timmerman, M.E. and Kiers, H.A.L., Four Simultaneous Component Models of multivariate time series from more than one subject to model intraindividual and interindividual differences. *Psychometrika*, 2003; **86**: 105-122

(5) Jackson, J.E., *A User's Guide to Principal Components*. Wiley-Interscience, New York, 1991

(6) Jolliffe, I.T., *Principal Component Analysis*. Springer-Verlag, New York, 2002

(7) Jansen, J.J., Hoefsloot, H.C.J., van der Greef, J., Timmerman, M.E. and Smilde, A.K., Multilevel Component Analysis of time-resolved metabolomics data. *Analytica Chimica Acta*, 2005; **530**: 173-183

(8) Pearce, S.C. and Holland, D.A., Some applications of multivariate methods in botany. *Appl. Stat.*, 1960; **9**: 1-7

(9) Jeffers, J.N.R., Principal component analysis of designed experiments. *The Statistician*, 1962; **12**: 230-242

(10) Mardia, K.V., Kent, J. and Bibby, J.M., *Multivariate Analysis*. Academic Press, London, 1979

(11) Stahle, L. and Wold, S., Multivariate Analysis of Variance (MANOVA). *Chemom. Intell. Lab. Syst.*, 1990; **9**: 127-141

(12) Martens, H. and Martens, M., *Multivariate Analysis of Quality*. John Wiley & Sons Ltd., Chichester, 2001

(13) Bratchell, N., Multivariate response surface modeling by Principal Component Analysis. *Journal of Chemometrics*, 1998; **3**: 579-588

(14) Hox, J.J., An introduction to Structural Equation Modeling. *Family Science Review*, 1998; **11**: 354-373

(15) Timmerman, M.E., Multilevel Component Analysis. *British Journal of Mathematical and Statistical Psychology*, 2005; In Press

(16) Jansen, J.J., Hoefsloot, H.C.J., Boelens, H.F.M., van der Greef, J. and Smilde, A.K., Analysis of longitudinal metabolomics data. *Bioinformatics*, 2004; **20**: 2438-2446

(17) Kiers, H.A.L. and Ten Berge, J.M.F., Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure. *British Journal of Mathematical and Statistical Psychology*, 1994; **47**: 109-126

(18) Ten Berge, J.M.F., Kiers, H.A.L. and Van der Stel, V., Simultaneous Component Analysis. *Statistica Applicata*, 1992; **4**: 377-392

(19) Jansen, J.J., Hoefsloot, H.C.J., Timmerman, M.E., van der Greef, J. and Smilde, A.K., Visualizing homeostatic capacity. Submitted;

(20) Keun, H.C., Ebbels, T.M., Bollard, M.E., Beckonert, O., Antti, H., Holmes, E., Lindon, J.C. and Nicholson, J.K., Geometric trajectory analysis of metabolic responses to toxcicity can define treatment specific profiles. 2004; **17**: 579-587

(21) Van den Brink, P.J. and Ter Braak, C.J.F., Principal Response Curves: Analysis of time-dependent multivariate responses of biological community to stress. *Environmental Toxicology and Chemistry*, 1997; **18**: 138-148

(22) Bro, R., PARAFAC: A Tutorial. *Chemom. Intell. Lab. Syst.*, 1997; **38**: 149-171

(23) Harshman, R.A., Foundations for the PARAFAC procedure: Model and conditions for an 'explanatory' multi-mode factor analysis. *UCLA working papers in phonetics*, 1970; **16**: 1-84

(24) Carroll, J.D. and Chang, J.J., Analysis of Individual Differences in Multidimensional scaling via an N-way generalization of "Eckart-Young" Decomposition. *Psychometrika*, 1970; **35**: 283-319

(25) Goodman, L.A. and Haberman, S.J., The analysis of nonadditivity in two-way analysis of variance. *Journal of the Neurological Sciences*, 1990; **85**: 139-145

(26) Smilde, A.K., Bro, R. and Geladi, P., *Multi-way Analysis: Applications in the Chemical Sciences*. John Wiley & Sons, New York, 2004

(27) Wold, S., Kettaneh, N., Fridèn, H. and Holmberg, A., Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemom. Intell. Lab. Syst.*, 1998; **44**: 331-340

(28) Barker, M., Rayens, W., Partial least squares for discrimination. *Journal of Chemometrics*, 2003; **17**: 166-173

(29) Carroll, J.D., Pruzansky, S. and Kruskal, J.B., CANDELINC: A general approach to multidimensional analysis of many-way data arrays with linear constraints on parameters. *Psychometrika*, 1980; **45**: 3-24

7 Future Work

None of the work in this thesis is ‘finished’. As in every scientific endeavor, each step towards an answer returns a whole array of new questions that demand an answer. This ‘future work’ section is subdivided into two sections, where the first deals with Weighted PCA and the second with ASCA.

7.1 WPCA

An important aspect of WPCA that is not yet developed is the selection of the number of model components. The total variation in the data can be separated into a contribution of *biologically* induced variation occurring from the subjects under investigation and *instrumentally* induced variation, caused by the measurement of the system. in Chapter 2, the WPCA weights are determined based on the instrumentally induced variation. Since the model residuals are weighted with this information, they should be representative of the instrumentally induced error. Consequently sufficient components should be chosen for the model such that the model mainly contains the (larger) biological variation and the residuals contain the (smaller) instrumental variation.

Unfortunately, certainly in problems of high-dimensional biology like metabolomics the ‘true’ underlying dimensionality of a problem is often not known (or even non-existing). Therefore alternative methods to determine the required number of components for a WPCA model need to be developed.

Another point of improvement for WPCA is the algorithm. The PCAW algorithm presented in Chapter 2 uses majorization iterations that are notoriously time-consuming. The MILES algorithm also uses this type of iterations. The alternating regression approach chosen for the MLPCA algorithm generally converges much faster. Modifying the MLPCA algorithm to include the comprehensive offset estimation described in the chapter will greatly increase the computational efficiency of Weighted PCA.

Already several applications of WPCA other than the analysis of data with a homoscedastic error were mentioned in Chapter 2 (autoscaling, data with missing values). The applicability of WPCA can be extended by devising weighting schemes to introduce other types of *a priori* information into a data analysis.

7.2 ASCA

Although in the final chapter of the thesis a framework for the ASCA model is presented, work in the method development of ASCA is far from complete.

One of the specific problems for which there is no answer yet is scaling. Specifically autoscaling poses a problem. It is performed in PCA to give every variable an equal variance and thereby an equal importance in the model. In more technical terms, PCA on unscaled data models the covariance matrix of the data while PCA on the autoscaled data models its correlation matrix.

In ASCA, there are two possibilities for scaling that will be illustrated using the decomposition of the data matrix $\mathbf{X}$ into different contributions, e.g. $\mathbf{X_a}$, $\mathbf{X_b}$ and $\mathbf{X_{(ab)}}$ in Chapter 6. The two available options for scaling are:

1. First autoscaling $\mathbf{X}$ and subsequently decomposing it into $\mathbf{X_a}$, $\mathbf{X_b}$ and $\mathbf{X_{(ab)}}$
2. First decomposing $\mathbf{X}$ and subsequently individually scaling $\mathbf{X_a}$, $\mathbf{X_b}$ and $\mathbf{X_{(ab)}}$

Both methods have drawbacks: When option 1 is used, the SCA models of $\mathbf{X_a}$, $\mathbf{X_b}$ and $\mathbf{X_{(ab)}}$ will not describe the correlation between the variables of each contribution, because in $\mathbf{X_a}$, $\mathbf{X_b}$ and $\mathbf{X_{(ab)}}$ the individual variables do not have a variance equal to one. When option 2 is used, the ASCA model does focus on the correlation between the variables within each contribution. However, the individually autoscaled matrices $\mathbf{X_a}$, $\mathbf{X_b}$ and $\mathbf{X_{(ab)}}$ do not add up to the autoscaled matrix $\mathbf{X}$ (disregarding the mean $\mathbf{1m}^T$). Therefore the SCA submodels of autoscaled matrices $\mathbf{X_a}$, $\mathbf{X_b}$ and $\mathbf{X_{(ab)}}$ and their residuals do also not add up to autoscaled $\mathbf{X}$ and therefore when option 2 is chosen for scaling the data, the model does not describe the original data anymore.

Clearly both options for scaling each have their own problems and scaling in ASCA requires further research. A possible solution might lie in scaling back the SCA submodels of autoscaled $\mathbf{X_a}$, $\mathbf{X_b}$ and $\mathbf{X_{(ab)}}$ such that they describe the original data matrix $\mathbf{X}$.

Another topic that remains subject of debate is the selection of an ANOVA model for metabolomics datasets such as the guinea pig dataset in Chapter 4 and the bromobenzene dataset in Chapter 5.

The purpose of ANOVA (and ASCA) is disentangling all contributions to the variation in a dataset. However, in the model given in Chapters 4 and 5, the variation of factor 'treatment group' (indicated by β) is not separated from the interaction between 'treatment group' and 'time point': both design parameters are confounded within one ANOVA term, which is indicated by $(\alpha\beta)$ in these chapters.

The rationale behind merging these two contributions is that the meaning of the factor 'treatment group' is not easy to interpret by itself. This contribution is an 'effect of each treatment averaged over all time-points', while the focus of the experiment is on the *evolving* behavior of the urine composition, such that this factor 'treatment group' is not primarily of interest.

The separation of the variation into a contribution equal for all animals, one for each treatment group and one for each individual as described in Chapters 4 and 5 leads to model parameters whose meaning is very intuitive, even though multiple contributions to the variation in the data are confounded within the treatment group-specific submodel. In ANOVA there is no dimension reduction and this problem does not exist, as described in paragraph 6.2.3.1.

A topic that is mostly untouched in this thesis is the statistical background of the ASCA method. The ASCA model requires a statistical background for two purposes. The quality of the scores and loadings fitted for the submodels has to be validated. Currently there is no method available to do this: work should be done in the development of methods for the determination of confidence intervals for the fitted model parameters (e.g. bootstrapping and other resampling methods).

Also the significance of the factors and interactions in the experimental design should be tested, analogous to significance testing in ANOVA. At present there is no method available for significance testing in ASCA. Significance testing in ASCA can greatly enhance its functionality.

Other promising work lies in the merger of ASCA with other data analysis methods. In Chapter 6 the use of multiway and regression models as ASCA submodels was already proposed. The SCA submodels of ASCA can also be replaced by Weighted PCA models. However, the choice of WPCA weights for the different submodels remains a subject of further research. Instead of other data analysis methods, also additionally constrained component models (e.g. in which explicit time-constraints are imposed for kinetics, smoothness etc.) can be used in ASCA, as long as the column spaces of the scores of the different submodels remain orthogonal.

Finally, the application of ASCA in various fields of data analysis (other than psychometrics and metabolomics) will be very challenging. Only the application of ASCA to a broad spectrum of scientific problems will show its added value with respect to the data analysis tools that are already available.

The use of other grey models for the analysis of time-resolved metabolomics data forms a promising subject of novel research. Aside from WPCA that uses information about the error and ASCA that uses the relationship between samples, another very promising model for metabolomics data analysis is the Network Component Analysis model (NCA) (1). In NCA, a partially known network structure is used as *a priori* information. Thereby the importance of the metabolites in the remainder of the metabolic network can be determined from the model.

7.3 References

(1) Liao, J.C., Boscolo, R., Yang, Y.L., Tran, L.M., Sabatti, C. and Roychowdhury, V.P., Network component analysis: reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences*, 2003; **100**: 1552-1557

8 Samenvatting

In dit proefschrift worden twee multivariate data-analyse methoden beschreven: Weighted PCA (WPCA) en ANOVA-Simultaneous Component Analysis (ASCA). Het gebruik van deze methoden wordt gedemonstreerd aan de hand van verschillende experimenten op het gebied van metabolomics.

Metabolomics is het vakgebied in de systeembiologie dat zich bezig houdt met het metabolisme van een organisme. Metabolomics wordt ondermeer vaak toegepast in farmacologische en toxicologische experimenten. In deze context, waarbij meestal modelorganismen als muizen, ratten of cavia's worden gebruikt, worden er lichaamsvloeistoffen van deze dieren afgenomen. Een vaak in metabolomics gebruikte lichaamsvloeistof is urine, aangezien het non-invasief en in redelijk grote hoeveelheden verkregen kan worden. De chemische (metaboliet)samenstelling van de verkregen urinesamples kan worden geanalyseerd met geavanceerde analytisch chemische technieken. Een vaak gebruikte techniek hiervoor is ^{1}H-Nuclear Magnetic Resonance (NMR) spectroscopie. Ook al is NMR spectroscopie niet erg gevoelig, waardoor slechts de metabolieten met de hoogste concentraties geanalyseerd kunnen worden, de zeer beperkte monstervoorbewerking maakt het een aantrekkelijke techniek voor het verkrijgen van een globaal overzicht van het metabolisme.

Lichaamsvloeistoffen als urine hebben een erg gecompliceerde chemische samenstelling en bovendien is de concentratie van nagenoeg elke metaboliet afhankelijk van de concentraties van andere metabolieten. Om uit deze grote hoeveelheid informatie de relevante en interessante fenomenen te extraheren worden multivariate data-analyse methoden gebruikt. Deze methoden geven een versimpeld en daardoor interpreteerbaar beeld van een dergelijke gecompliceerde metabolomics dataset.

Een vaak gebruikte methode voor de analyse van metabolomics datasets is Principal Component Analysis (PCA): een krachtige methode die de relatie tussen de concentraties van verschillende metabolieten gebruikt om de grootste variatie in een metabolomics dataset weer te geven. Echter, PCA heeft ook enige tekortkomingen die door het gebruik van de methoden beschreven in dit proefschrift verholpen kunnen worden.

In hoofdstuk 2 van het proefschrift wordt de analyse van een metabolomics dataset met WPCA geïllustreerd. In dit hoofdstuk laat een analyse van de fout in de data geïntroduceerd door de NMR-metingen zien, dat deze afhankelijk is van de grootte van het signaal (dus van de concentratie van een metaboliet): de meetfout is 'heteroscedastisch'. Wanneer PCA voor de analyse van deze data wordt gebruikt, zijn de metabolieten met de hoogste concentraties relatief het meest belangrijk. Echter, de concentratiemetingen van deze metabolieten hebben dus ook de grootste fout. WPCA weegt iedere meting in de dataset met de eerder bepaalde meetfout. Deze methode maakt daardoor de metingen met een grote meetfout relatief minder belangrijk dan de metingen met een kleine meetfout. Hierdoor heeft de heteroscedastische meetfout geen invloed op de resultaten van het WPCA model.

In de andere hoofdstukken wordt een andere tekortkoming van PCA onder de loep genomen: deze methode kijkt namelijk naar de variatie in een dataset, maar het houdt geen rekening met de origine van deze variatie. Voor een goed begrip van de fenomenen die ten grondslag liggen aan de variatie in zo een dataset is het echter van groot belang dat de totale variatie in een dataset onderscheiden wordt naar verschillende types die toegeschreven kunnen worden aan verschillende origines.

Het experiment dat besproken wordt in hoofdstuk 3 bestaat uit metingen van de normale variatie in de urinecompositie van 10 verschillende apen. De metabolietcompositie van de urine zal variëren in de tijd. Er zullen echter ook verschillen zijn tussen de urinecompositie van de verschillende apen die constant zijn in de tijd. In dit hoofdstuk wordt Multilevel Simultaneous Component Analysis (MSCA) geïntroduceerd: een methode die onderscheid maakt tussen deze types variatie en een interpreteerbaar beeld geeft van beide. Met het gebruik van MSCA kunnen metabolieten die alleen maar in de tijd of alleen maar tussen de apen variëren geïdentificeerd worden, zodat het begrip van de chemische achtergrond van de variatie in metabolietcompositie duidelijk verbeterd wordt.

MSCA is een specifieke versie van ASCA, dat in de volgende hoofdstukken besproken wordt.

In hoofdstuk 4 wordt het onderscheid naar verschillende origines van variatie in een metabolomics experiment uitgebreid: de dataset die in dit hoofdstuk geanalyseerd wordt is afkomstig van een experiment waarin een bepaald type genetisch gemodificeerde cavia's osteoarthritis, een ziekte aan de gewrichten, krijgt. Door middel van het toedienen van verschillende doses Vitamine C aan deze cavia's wordt bekeken of het een invloed op de ontwikkeling van osteoarthritis in de tijd heeft. In dit experiment is er dus een bijdrage in de variatie die gelijk is voor alle cavia's, er is een bijdrage in de variatie die verschilt tussen de verschillende dosisgroepen maar gelijk is voor alle cavia's binnen een dosisgroep en er is een bijdrage die uniek is voor iedere cavia. Met ASCA wordt de totale variatie in deze dataset opgedeeld in deze verschillende bijdragen en iedere bijdrage wordt onafhankelijk geanalyseerd. Uit deze analyse blijkt dat er geen verschil is tussen de verschillende doses Vitamine C en dat deze stof geen invloed heeft op de ontwikkeling van osteoarthritis in de tijd.

In hoofdstuk 5 wordt ASCA gebruikt voor de analyse van de 'homeostatic capacity': een begrip uit de systeembiologie dat aangeeft of een bepaald dier binnen een bepaald tijdsbestek in staat is om een verstoring van het metabolisme (bijvoorbeeld geïnduceerd door het toedienen van een vergif) te verwerken en terug te komen tot het 'normale' niveau van het metabolisme. Dit wordt geïllustreerd aan de hand van een experiment met ratten waaraan verschillende hoeveelheden broombenzeen, een stof die ontstekingen veroorzaakt in de lever, toegediend wordt. Met behulp van ASCA blijkt de 'homeostatic capacity' van ratten voor broombenzeen goed vastgesteld te kunnen worden.

In hoofdstuk 6 wordt de theoretische achtergrond van ASCA uitgelegd. Met de informatie in dit hoofdstuk kan de onderzoeker de data verkregen uit zijn eigen experiment analyseren met behulp van een zelf geconstrueerd ASCA model, gebaseerd op de gedefinieerde types van variatie in de data. Allereerst wordt ASCA uitgelegd, daarna worden de eigenschappen van ASCA vergeleken met die van PCA en ANOVA (de twee bouwstenen van ASCA). Er wordt aangegeven hoe de kwaliteit van een ASCA model voor een bepaalde dataset bepaald kan worden en het gebruik van ASCA voor de analyse van de data van verschillende types experimenten wordt geïllustreerd. Verder wordt ASCA vergeleken met verschillende gerelateerde multivariate data-analyse methoden. Als laatste wordt ASCA in een breder perspectief geplaatst: ANOVA kan niet alleen met PCA gecombineerd worden tot ASCA, maar ook met andere data-analyse methoden (zoals de multiway-methode PARAFAC of de multivariate regressiemethode PLS) om andere multivariate data-analysemethoden te maken waarin verschillende contributies gedefinieerd worden.

9 Summary

In this thesis two different multivariate data-analysis methods are described: Weighted PCA (WPCA) and ANOVA-Simultaneous Component Analysis (ASCA). The use of these methods is demonstrated using different experiments from the field of metabolomics.

Metabolomics is the field of Systems Biology that deals with the metabolism of an organism. Metabolomics is a widely used method in pharmacology and toxicology. In this context, in which most often laboratory model-animals like rats, mice or guinea pigs are used, body fluids of these animals are collected. A body fluid that is often used in metabolomics is urine, since it can be collected non-invasively and in reasonably large amounts. The chemical (metabolite) composition of the collected urine samples can be analyzed by advanced analytical chemical methods. An often used technique for this is ^{1}H-Nuclear Magnetic Resonance (NMR) spectroscopy. Although NMR spectroscopy is not very sensitive, such that only the metabolites with relatively large concentrations can be analyzed, the very simple sample preprocessing makes it a very attractive technique for obtaining a global overview of the metabolism.

Body fluids like urine have a very complex chemical composition; furthermore the concentration of basically every metabolite dependent on the concentrations of other metabolites. To extract the relevant and interesting phenomena from this large amount of information, multivariate data-analysis methods are used. These methods give a simplified and therefore interpretable view on these complicated metabolomics datasets.

An often used method for the analysis of metabolomics datasets is Principal Component Analysis (PCA): a powerful method that uses the relationship between the concentrations of different metabolites to describe the largest variation in a metabolomics dataset. However, PCA also has some drawbacks that can be countered by using the alternative methods described in this thesis.

In chapter 2 of the thesis the analysis of a metabolomics dataset using WPCA is illustrated. In this chapter the analysis of the error introduced by performing the measurements using NMR spectroscopy shows that it is dependent on the size of the NMR signal (and therefore on the concentration of a metabolite): the measurement error is ‘heteroscedastic’. When PCA is used for the analysis of this data, the metabolites that have the highest concentration are relatively most important. However, the measurement of the concentration of these metabolites also has the largest measurement error. WPCA weighs each measurement in the dataset with the measurement error that was determined earlier. Thereby this method makes the measurements with the highest error less important and the measurements with the smallest error more important. This removes the influence of the heteroscedastic measurement error on the results of the WPCA model.

In the other chapters of the thesis the focus is on another shortcoming of PCA: this method focuses on the variation in a dataset, but it does not distinguish different origins of the variation. For a thorough understanding of the phenomena underlying this variation it is important that the total variation in a dataset is distinguished into different contributions that can be assigned to different origins of variation.

The experiment that is discussed in chapter 3 consists of measurements of the normal variation in time of the urine composition collected from 10 different monkeys. The metabolite composition of the urine will vary in time, but there will be also differences between the urine compositions of the different monkeys. In this chapter Multilevel Simultaneous Component Analysis (MSCA) is introduced: a method that distinguishes between these types of variation and gives an interpretable view on both. By using MSCA metabolites that vary only between the monkeys and not in time (or vice versa) can be identified, to increase the understanding of the chemical background of the variation in urine composition. MSCA is a specific version of ASCA: the method that is described in the subsequent chapters.

In chapter 4 the distinction between different origins of variation in a metabolomics experiment is extended: the dataset described in this chapter is collected from a metabolomics experiment in which a specific type of genetically modified guinea pigs develops osteoarthritis: a disease of the joints. By administering different doses of Vitamin C, the influence of this compound on the development of the disease is determined. In this experiment a contribution to the variation that is equal for all guinea pigs, a contribution that is equal for all guinea pigs within a dose group and a contribution that is unique to each guinea pig can be distinguished. Using ASCA the total variation in the dataset is divided into these contributions, such that each contribution can be individually analyzed. From this analysis the conclusion is drawn that there are no differences between the different Vitamin C dose groups, such that the compound has no influence on the development of osteoarthritis.

In chapter 5 ASCA is used for the analysis of the ‘homeostatic capacity’: an entity from systems biology that indicates the ability of a specific animal to regain a ‘normal’ metabolism (homeostasis) within a predefined time-span after a perturbation of the metabolism, for example induced by a toxic compound. This is illustrated using an experiment performed on rats. To these rats different doses of bromobenzene, a liver toxic compound, is administered. Using ASCA the homestatic capacity of rats for bromobenzene can be well determined.

In chapter 6 the theoretical background of ASCA is explained. Using the information in this chapter, the researcher can analyze his/her data using a self-constructed ASCA model, based on the defined contributions to the variation in the data. First the principles behind ASCA are explained; subsequently the properties of ASCA are compared to those of PCA and ANOVA (its building blocks). A method to determine the quality of a constructed ASCA model for describing a specific dataset is given and the use of ASCA for the analysis of data collected from different types of experiments is indicated. Also ASCA is compared to different related multivariate data-analysis methods. Finally, ASCA is put into a broader perspective: ANOVA can not only be combined with PCA to ASCA, but also with other multivariate data-analysis methods (like the multiway-method PARAFAC or the multivariate regression method PLS) to make other novel multivariate data-analysis methods based on multiple contributions to the total variation in a dataset.

10 Acknowledgments

First and foremost I would like to thank Huub and Age for helping me writing this thesis, as well as providing a very nice working environment these past years. Age's thoroughness combined with Huub's inventiveness and smart ideas have been very important for the successful finishing of this thesis and you have really taught me what being a scientist is all about.

Furthermore I would like to thank the other colleagues in the group.
First of all Hans: thank you for all the help in the beginning of my Ph. D. research. You have really taught me a lot about MATLAB and also about keeping all my files tidily arranged (although there are still about 500 files named 'aap.m' on my hard disk).
Eric and Henk Jan: thank you for all the advice you have given me and especially thank you for all the friendship during the past four years: it really wouldn't have been as much fun without you! Also thanks: I will keep on making use of it. It will be hard for me to find other colleagues that are able to replace you.
Suzanne: I was very amazed that you popped up again as a colleague after 2 years and it has been a very nice time working with you! Now I will have to find someone else for the Metro puzzle, bugger!
Erik: Mange Tak for becoming a friend and providing a welcoming place everytime I went to Copenhagen.
Maikel: thank you for being such a nice roommate. I hope you will be able to get your M. Sc. as quickly as possible so you can go on to do interesting things!
Olja: thank you for all the nice times together! Good luck together with Hans and Tessa!
Also many thanks to the other colleagues: Johan, Robert, Susana, Jos and Panos, it has been very pleasant working with you!

During the past four years I have also collaborated with people outside the UvA Biosystems Data Analysis group.

Marieke: thank you for the nice collaboration the past years. Your comments have been very important to the work in this thesis and I’ve always found it amazing that I could collaborate with a psychologist in my research which is focused on biology, chemistry and medicine.

I would also like to thank the people I worked with at TNO Quality of Life in Zeist:
Jan: thank you for all the good advice about the articles we have written (and therefore about the contents of this thesis).
Robert-Jan: thank you for our discussions about metabolomics and about the guinea pig and the Bromobenzene data sets.
Jack: thank you for telling me about NMR spectroscopy!
Also to the other people I’ve encountered at TNO: Renger, Sabina, Carina, Uwe, Florian and Gerwin thank you!

Thank you to the commission members for reading my thesis. I hope it has been a pleasant read!

I also would like to thank my family for their support the past four years: Nanda, mama, papa, Bert en Jopie dankjewel! Het is altijd leuk geweest om jullie over de leuke dingen in het werk te vertellen maar ook goed om mijn frustraties bij jullie kwijt te kunnen als het eens niet zo ging als ik wilde!

En natuurlijk niet te vergeten: Dankjewel Nicole voor de koffie en het gejodel! Mijn achterstallige betalingen komen er aan!